

How Generative Language Models Shape Human Problem Solving: A Phenomenon-centered, Sociotechnical Perspective

Cumulative dissertation with the aim of achieving a doctoral degree at the
Faculty of Mathematics, Informatics, and Natural Sciences
Department of Informatics
Universität Hamburg

Submitted by

Lucas Memmert

2025

Hamburg, Germany

Evaluators

First Evaluator and Supervisor

Prof. Dr. Eva Bittner

Second Evaluator

Prof. Dr. Tilo Böhmann

Third Evaluator

Prof. Dr. Gerhard Schwabe

Chair of the Examination Committee

Prof. Dr. Tilo Böhmann

Deputy Chair of the Examination Committee

Prof. Dr. Frank Steinicke

Date of disputation

13.02.2026

Abstract

Motivation & research goal

Problem-solving is essential for knowledge work and innovation in organizations. Text-based generative large language models (GLMs) are being rapidly adopted across domains, offering potentially substantial performance improvements. Yet, GLMs can also introduce risks and negatively affect outcomes, creating a pressing practical need to better understand how humans can productively work with GLMs.

This need from practice ties into the academic discourse on human–AI collaboration, which examines how humans and AI jointly work on tasks. Specifically, scholars have called for research that explains how humans and AI systems can jointly achieve superior performance and that uncovers the underlying dynamics of such collaboration when solving problems. Due to their domain-agnostic generative capacity—the ability to create novel content in real-time across a broad range of topics—GLMs enable a new form of human–AI work. They can dynamically support open-ended problem-solving tasks at the content level, rather than supporting routine decision-making or facilitating at the meta-level. Challenging how humans can work with AI systems renders human–GLM problem-solving a theoretically interesting phenomenon. Accordingly, this dissertation seeks to explain how humans work with GLMs to solve problems and to uncover mechanisms and factors that shape performance and behavior.

Research design

To address this goal, the dissertation adopts a phenomenon-centered research approach, framing human–GLM problem-solving as a sociotechnical phenomenon. It follows a cumulative, multi-study design comprising three journal and eight conference publications. The empirical, multi-method approach combines exploratory and confirmatory methods in selected tasks and contexts. These include brainstorming, design science research concept development, coding, and professional knowledge work. Data collection includes qualitative and quantitative surveys, interviews, and a diary study, as well as interaction and usage data from controlled experiments utilizing GLM-based custom-built web applications and commercial, chat-based GLMs. The data analyses span qualitative content analysis, and descriptive and inferential statistics. They draw on theories on social loafing, cognitive stimulation, and cognitive load, building on decades of research on group work and individual creativity support systems.

Results

The studies show empirically that human–GLM dyads can achieve superior performance in creative idea generation, a key aspect of problem-solving. They further disentangle performance, revealing tensions between individual and dyadic performance as well as between perceived and measured outcomes.

The dissertation identifies four mechanisms across studies through which GLMs shape performance and human behavior: (1) Dynamic scaffolding describes how GLMs help individuals move efficiently through a problem by, for example, supporting operationalization or preventing getting stuck; (2) Problem- and solution space exploration captures how GLMs broaden or narrow the set of options considered; (3) Effort management sheds light on humans' shift from being sole creators of content to creating own and curating GLM content; and (4) Responsibility attribution describes how GLMs change users' felt responsibility. The dissertation also discusses influencing factors, including individual ability or contextual norms, and formulates tentative design implications based on the findings.

Contribution

The dissertation primarily contributes to two discourses in the literature: human–AI collaboration and individual creativity support systems. For human–AI collaboration, it offers an empirically grounded and sharpened understanding of human–GLM problem-solving, thereby enhancing concept clarity. Moreover, it provides explanatory knowledge by describing mechanisms and potential boundary conditions, an important foundation for design knowledge. The findings help to inform the discourse on the transferability of human group-level mechanisms to human–AI dyads, extending the social loafing construct beyond its original scope and suggesting how it may need to be adapted.

This dissertation extends the individual creativity support systems research stream by framing GLMs as a new type of individual creativity support system. It offers empirical evidence that human–GLM dyads can outperform individuals in idea generation, as well as how GLMs influence the diversity of outputs, challenging prior notions of how AI can support creative tasks. Lastly, it suggests that GLMs appear to have a different effect on humans compared to prior individual creativity support systems or human group members.

Limitations

Beyond the study-specific limitations discussed in the individual publications, the dissertation has several overarching limitations related to its scope and its research design and empirical approach, which limit generalization despite method and data triangulation. Moreover, its theorizing is limited, with the dissertation focusing on developing an in-depth explanatory understanding of the phenomenon rather than developing and evaluating prescriptive design knowledge. As such, the design implications remain tentative and require further empirical validation.

Future Research

The findings in this dissertation open up several avenues for future research. While GLMs can improve performance, they also alter human behavior—sometimes in undesirable ways. Future research should examine how tools can be designed to counteract such behavior and support users in situationally improving their allocation of effort between creating and curating. Moreover, extending the dissertation's findings, future research should

explore how more capable or more agentic AI, or the involvement of multiple AI systems or agents, shape felt responsibility or effort investment.

Professional knowledge workers appear to experience a “fill-up effect,” in which saved time due to working with the GLM is used to complete additional tasks. Aligning with calls for examining humanistic, not solely instrumental goals, future research should explore whether such patterns contribute to a “mechanization” of knowledge work and potential implications for worker well-being, particularly if knowledge work shifts from creation to curation. Finally, while the studies show that GLMs can support creative tasks, they did not investigate professional creative workflows involving extended, iterative projects. Future research should examine whether GLMs can support creative professionals in the creative part of their work, beyond merely enhancing efficiency.

Keywords

Human–AI Collaboration, Problem-Solving, Generative Large Language Models, Brainstorming, Effort, Performance

Kurzfassung

Motivation

Das Lösen von Problemen ist zentral für Wissensarbeitende und Innovation in Organisationen. Zunehmend werden hierfür domänenübergreifend textbasierte generative Sprachmodelle (engl. Generative Language Models, GLMs) eingesetzt, die erhebliche Verbesserungen bei der Aufgabenbearbeitung ermöglichen können. Gleichzeitig bergen GLMs jedoch Risiken und können die Ergebnisse beeinträchtigen. Aus der Praxis ergibt sich damit ein dringender Bedarf, besser zu verstehen, wie wir mit GLMs produktiv zusammenarbeiten können.

Dieser praktische Bedarf spiegelt sich auch im wissenschaftlichen Diskurs zur Mensch-KI-Zusammenarbeit oder -Kollaboration wider. Dieser untersucht, wie Menschen und KI-Systeme gemeinsam Aufgaben bearbeiten. Ziel ist es insbesondere zu erklären, wie Menschen und KI-Systeme gemeinsam eine bessere Leistung (engl. Performance) erzielen können und welche Dynamiken bei dieser Zusammenarbeit entstehen. Aufgrund ihrer domänenagnostischen generativen Fähigkeit – der Fähigkeit, in Echtzeit neuartige Inhalte über ein breites Themenspektrum hinweg zu erzeugen – ermöglichen GLMs eine neue Form der Mensch-KI-Zusammenarbeit. Sie können bei der Lösung offener Probleme (engl. open-ended problems) dynamisch auf inhaltlicher Ebene unterstützen, statt lediglich routinemäßige Entscheidungen oder nur auf der Metaebene zu unterstützen. Dies macht die Frage, wie Menschen und GLMs zusammenarbeiten, um Probleme zu lösen, auch theoretisch interessant. Vor diesem Hintergrund hat diese Dissertation das Ziel, zu erklären, wie Menschen mit GLMs Probleme lösen, und Mechanismen und Einflussfaktoren zu beschreiben, die die Ergebnisse und das menschliche Verhalten prägen.

Forschungsdesign

Hierfür verfolgt die Dissertation einen phänomenzentrierten Forschungsansatz und nimmt eine soziotechnische Perspektive auf das Mensch-GLM-Problemlösen (engl. human-GLM problem-solving) ein. Sie folgt einem kumulativen Multi-Studien-Design und umfasst drei Zeitschriften- und acht Konferenzpublikationen. Der empirische Multi-Methoden-Ansatz kombiniert explorative und konfirmatorische Methoden in ausgewählten Aufgaben und Kontexten, darunter Brainstorming, Konzeptentwicklung für Design-Science-Research-Studien, Programmieren sowie professionelle Wissensarbeitsaufgaben. Die Datenerhebung umfasst qualitative und quantitative Befragungen, Interviews und eine Tagebuchstudie sowie Interaktions- und Nutzungsdaten aus kontrollierten Experimenten mit GLM-basierten, eigens entwickelten Webanwendungen sowie kommerziellen, chatbasierten GLMs. Die Datenanalysen umfassen qualitative Inhaltsanalysen sowie deskriptive und inferenzstatistische Verfahren. Theoretisch stützt sich die Arbeit auf Konzepte wie Social Loafing, kognitive Stimulation und kognitive Belastung und baut auf jahrzehntelanger Forschung zu Gruppenarbeit sowie zu Kreativitätsunterstützungssystemen für Individuen auf.

Ergebnisse

Die Studien zeigen empirisch, dass Dyaden aus Mensch und GLM bei der kreativen Ideengenerierung – einem zentralen Aspekt des Problemlösens – eine überlegene Leistung gegenüber nicht durch GLM unterstützten Personen erzielen können. Gleichzeitig differenziert die Arbeit den Leistungsbegriff weiter aus und zeigt Spannungsfelder zwischen individueller und dyadischer Leistung sowie zwischen wahrgenommener und gemessener Leistung auf.

Über die Studien hinweg identifiziert die Dissertation vier Mechanismen, durch die GLMs Leistung und menschliches Verhalten beeinflussen: (1) Dynamic Scaffolding, also die Bereitstellung dynamischer Hilfskonstruktionen, beschreibt, wie GLMs Individuen dabei unterstützen, Probleme effizient abzuarbeiten, etwa durch die Unterstützung bei der Operationalisierung abstrakter Ideen oder durch das Verhindern mentaler Blockaden; (2) Exploration des Problem- und Lösungsraums beschreibt, wie GLMs die in Betracht gezogenen Optionen erweitern oder verengen kann; (3) Aufwandsmanagement (engl. Effort Management) beleuchtet die Verschiebung des Fokus der menschlichen Tätigkeit von der alleinigen Erstellung von Inhalten hin zur Erstellung eigener Inhalte und der Kuratierung von GLM-generierten Inhalten; und (4) Verantwortungszuweisung (eng. Responsibility Attribution) beschreibt, wie GLMs das subjektiv empfundene Verantwortungsgefühl der Nutzenden verändern. Darüber hinaus diskutiert die Dissertation Einflussfaktoren wie individuelle Fähigkeiten oder kontextuelle Normen und leitet auf Grundlage der Ergebnisse vorläufige Designimplikationen ab.

Beitrag

Die Dissertation leistet Beiträge insbesondere zu zwei Forschungssträngen: der Mensch-KI-Zusammenarbeit und den Kreativitätsunterstützungssystemen für Individuen (engl. individual creativity support systems). Im Bereich der Mensch-KI-Zusammenarbeit bietet sie ein empirisch fundiertes und geschärftes Verständnis des Problemlösens durch Mensch und GLM und trägt damit zur Definition zentraler Konzepte bei. Darüber hinaus liefert sie erklärendes Wissen, indem sie zugrunde liegende Mechanismen und potenzielle Randbedingungen beschreibt, was eine wichtige Grundlage für die Entwicklung von Gestaltungswissen darstellt. Die Ergebnisse tragen zudem zum Diskurs über die Übertragbarkeit von Mechanismen aus der Forschung zu menschlichen Gruppen auf Mensch-KI-Dyaden bei und zeigen auf, wie das Social-Loafing-Konzept über seinen ursprünglichen Anwendungsbereich hinaus angepasst werden könnte.

Zudem erweitert die Dissertation die Forschung zu Kreativitätsunterstützungssystemen für Individuen, indem sie GLMs als einen neuen Typ dieser Systeme konzeptualisiert. Sie liefert empirische Evidenz dafür, dass Mensch-GLM-Dyaden Individuen bei der Ideengenerierung übertreffen können, und zeigt auf, wie GLMs die Diversität der Ergebnisse beeinflussen, wodurch bestehende Annahmen zur Unterstützung von Kreativität durch KI hinterfragt werden. Schließlich legt die Arbeit nahe, dass GLMs eine andere Wirkung auf

Menschen entfalten als frühere individuelle Kreativitätsunterstützungssysteme oder als menschliche Gruppenmitglieder.

Limitationen

Über die studienspezifischen Limitationen der einzelnen Publikationen hinaus weist die Dissertation mehrere übergreifende Einschränkungen auf, die sich aus ihrer thematischen Abgrenzung sowie dem Forschungsdesign und dem empirischen Ansatz ergeben, wodurch die Generalisierbarkeit der Ergebnisse trotz Methoden- und Datentriangulation begrenzt bleibt. Zudem ist der theoretische Beitrag der Arbeit in seiner Art beschränkt, da der Fokus auf der Entwicklung eines vertieften erklärenden Verständnisses des Phänomens liegt und nicht auf der Entwicklung und Evaluation präskriptiven Gestaltungswissens. Entsprechend bleiben die abgeleiteten Designimplikationen vorläufig und bedürfen weiterer empirischer Validierung.

Ausblick

Aus den Ergebnissen dieser Dissertation ergeben sich mehrere Ansatzpunkte für zukünftige Forschung. Wenngleich GLMs die Leistung steigern können, verändern sie auch menschliches Verhalten – teilweise in unerwünschter Weise. Zukünftige Forschung sollte untersuchen, wie Systeme gestaltet werden können, um solchen Effekten entgegenzuwirken und Nutzende dabei zu unterstützen, ihre Aufwandsverteilung zwischen dem Erstellen und dem Kuratieren von Inhalten situationsabhängig zu verbessern. Darüber hinaus sollte erforscht werden, wie leistungsfähigere oder autonomere KI-Systeme sowie der Einsatz mehrerer KI-Systeme oder Agenten das empfundene Verantwortungsgefühl oder den Aufwand bzw. das Engagement beeinflussen.

Laut unserer Studie erleben Wissensarbeitende etwas, das sich als Auffülleffekt (engl. fill-up effect) beschreiben lässt, wonach sie eingesparte Zeit durch die Zusammenarbeit mit GLMs nutzen, um zusätzliche Aufgaben zu erledigen. In Anlehnung an Forderungen, neben instrumentellen auch humanistische Ziele stärker zu berücksichtigen, sollte zukünftige Forschung untersuchen, ob solche Muster zu einer „Mechanisierung“ von Wissensarbeit beitragen und welche Auswirkungen dies auf das Wohlbefinden von Beschäftigten haben könnte – insbesondere, wenn sich Wissensarbeit zunehmend von der Erstellung hin zur Kuratierung von Inhalten verlagert. Zudem haben die vorliegenden Studien zwar gezeigt, dass GLMs kreative Aufgaben unterstützen können, jedoch keine professionellen kreativen Arbeitsprozesse mit längeren, iterativen Projekten untersucht. Zukünftige Forschung sollte daher analysieren, ob und wie GLMs Kreativarbeitende auch im kreativen Teil ihrer Arbeit unterstützen können – über reine Effizienzsteigerungen hinaus.

Schlagwörter

Mensch–KI-Zusammenarbeit, Problemlösen, Generative Large Language Models, Brainstorming, Aufwand, Leistung

Table of Contents

I. List of figures	XI
II. List of tables	XIII
III. List of abbreviations	XV
1 Introduction	1
1.1 Motivation and problems	1
1.2 Research goal and research questions	3
1.3 Thesis outline	5
2 Background	7
2.1 Human–GLM problem-solving.....	7
2.1.1 Human–AI collaboration.....	7
2.1.2 Sociotechnical perspective as a lens.....	8
2.1.3 Problem-solving as a type of situated task	9
2.1.4 Generative language models as enabling technology.....	11
2.2 Creative idea generation.....	12
2.2.1 Brainstorming.....	12
2.2.2 Individual creativity support systems.....	13
2.3 Mechanisms in (human) groups	13
3 Research Design	15
3.1 Phenomenon-centered research.....	15
3.2 Sociotechnical framing of the phenomenon.....	16
3.3 Empirical, multi-method approach.....	16
4 Publications	19
4.1 Overview	19
4.2 Key information on the publications.....	19
4.2.1 Paper 1: Memmert and Bittner (2022)	21
4.2.2 Paper 2: Memmert and Tavanapour (2023)	22
4.2.3 Paper 3: Memmert, Cvetkovic, and Bittner (2024).....	23
4.2.4 Paper 4: Memmert (2024)	24
4.2.5 Paper 5: Memmert and Bittner (2024)	25
4.2.6 Paper 6: Memmert, Cvetkovic, et al. (2025)	26
4.2.7 Paper 7: Memmert, Mies, and Bittner (2024)	27
4.2.8 Paper 8: Memmert, Tavanapour, and Bittner (2023)	28
4.2.9 Paper 9: Memmert, Cvetkovic, and Bittner (2023).....	29
4.2.10 Paper 10: Memmert, Borchers, et al. (2025)	30
4.2.11 Paper 11: Memmert, Soroko, and Bittner (2025).....	31
5 Integrative discussion	33
5.1 Untangling human–GLM task performance (RQ 1)	33
5.1.1 Facets of performance	33

5.1.2	GLM's direct and indirect performance influence.....	34
5.1.3	Collective and individual performance	35
5.1.4	Subjective vs. quantitatively measured performance	37
5.1.5	Summary of performance implications	38
5.2	Mechanisms: Effects on behavior and problem-solving performance (RQ 2)..	39
5.2.1	Dynamic scaffolding	39
5.2.2	Problem- and solution space exploration	42
5.2.3	Effort management.....	44
5.2.4	Responsibility attribution.....	48
5.3	Influencing factors (RQ 3)	51
5.3.1	Individual: Task-related ability	51
5.3.2	Situated task: Contextual norms and task structure.....	53
5.3.3	Technology: Intricacies of GLMs	55
5.4	Implications for human–GLM design (RQ 4).....	57
5.4.1	GLM-based adaptive support.....	57
5.4.2	Embedded prompt design.....	58
5.4.3	Tentative implications for design.....	59
6	Theoretical contribution	61
6.1	Contributions to the conversation in (IS) literature.....	61
6.1.1	Human–AI collaboration for problem-solving.....	61
6.1.2	Superior performance.....	62
6.1.3	Performance mechanisms and human behavior	63
6.1.4	Contribution to individual creativity support systems	65
6.2	Summary of theoretical contributions	67
6.3	Theoretical discussions on an emerging phenomenon	68
7	Practical implications	71
7.1	Managers and organizations.....	71
7.2	Tool designers	72
7.3	Knowledge workers	73
8	Limitations and outlook	75
8.1	Limitations to scope	75
8.2	Limitations in research design and empirical approach	76
8.3	Limitations of theorizing.....	77
8.4	Directions for future research.....	77
8.4.1	Human–AI constellations beyond dyads.....	78
8.4.2	AI autonomy and capability	78
8.4.3	Formation of effort management.....	78
9	Conclusion	81
	References	83

10 Publication 1: Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts.....	95
10.1 Introduction.....	95
10.2 Background.....	97
10.2.1 (Human-AI) collaboration.....	97
10.2.2 Complex problem solving.....	99
10.3 Methodology.....	102
10.4 Results.....	103
10.4.1 Research contexts of human-AI collaboration (RQ 1).....	103
10.4.2 Complex problem solving (RQ 2).....	107
10.5 Limitations.....	110
10.6 Conclusion and outlook.....	110
10.7 References.....	111
11 Publication 2: Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI	117
11.1 Introduction.....	117
11.2 Background.....	118
11.2.1 Brainstorming.....	118
11.2.2 Human-AI collaboration in brainstorming.....	119
11.2.3 Content-level support through generative AI.....	121
11.3 Method.....	122
11.3.1 Data collection.....	122
11.3.2 Data analysis.....	124
11.4 Results.....	125
11.4.1 Overview: working with the AI system.....	125
11.4.2 Cognitive stimulation/synergy.....	125
11.4.3 Free riding.....	126
11.4.4 Observational learning and negative productivity matching.....	129
11.4.5 Cognitive inertia and cognitive interference.....	129
11.4.6 Group setting preference: alone, with AI, or with another human.....	130
11.5 Discussion and future research.....	131
11.5.1 Process effects in human-AI groups.....	131
11.5.2 Generative AI for collaboration on content-level when solving problems	133
11.6 Conclusion.....	134
11.7 References.....	135
12 Publication 3: The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3	141
12.1 Introduction.....	141

12.2	Background	143
12.2.1	Idea generation support	143
12.2.2	Generative language models (GLM).....	143
12.3	Method	144
12.3.1	Prompt template development.....	144
12.3.2	Idea generation	148
12.3.3	Idea evaluation	148
12.3.4	Data analysis	149
12.4	Results.....	149
12.5	Discussion	151
12.5.1	Answer to the research question.....	151
12.5.2	Contribution to theory	152
12.5.3	Implications for practice	153
12.5.4	Limitations & outlook	153
12.6	Conclusion	155
12.7	References	155
13	Publication 4: Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects ..	161
13.1	Introduction	161
13.2	Background	163
13.2.1	Brainstorming Group Effects	163
13.2.2	Cognitive Forcing.....	164
13.2.3	Hypothesis Development	164
13.3	Method	166
13.3.1	Pre-study	166
13.3.2	Adaptation for planned study	166
13.4	Results so far	167
13.5	Discussion and Implications	168
13.6	Conclusion and Outlook.....	170
13.7	References	170
14	Publication 5: Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration	173
14.1	Introduction	173
14.2	Background	175
14.2.1	Brainstorming.....	175
14.2.2	Human-AI brainstorming	176
14.3	Method	178
14.3.1	Procedure and participants	178
14.3.2	Data collection instrument	178
14.3.3	Data preparation and analysis	179

14.4	Results.....	180
14.4.1	Breadth of exploration.....	181
14.4.2	Driving category exploration	182
14.4.3	Subjective perception.....	183
14.5	Discussion	184
14.5.1	Answer to the research question.....	184
14.5.2	Contribution and implications.....	185
14.5.3	Limitations	186
14.6	Conclusion	187
14.7	References.....	187
15	Publication 6: Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load	191
15.1	Introduction.....	191
15.2	Background	195
15.2.1	Brainstorming Performance and Determinants	195
15.2.2	Towards Human–AI Collaboration for Brainstorming	196
15.2.3	Hypotheses Development.....	198
15.3	Method	203
15.3.1	Experimental Design and Procedure.....	203
15.3.2	Experimental Treatments	203
15.3.3	Material: GPT-3.5-based Brainstorming Application and Use Process..	204
15.3.4	Participants.....	205
15.3.5	Measures	205
15.3.6	Data Collection, Preparation, and Analysis	207
15.4	Results.....	208
15.4.1	Hypothesis Set 1: Human-Level Performance	208
15.4.2	Hypothesis Set 2: Dyad-Level Performance	209
15.4.3	Hypothesis 3: Cognitive Load.....	210
15.4.4	Explorative Analysis of Performance-affecting Aspects	211
15.5	Discussion	215
15.5.1	General Discussion.....	215
15.5.2	Contributions to Literature and Theory.....	220
15.5.3	Implications for Practice	221
15.5.4	Limitations	222
15.5.5	Outlook.....	223
15.6	Conclusion	223
15.7	References.....	224
15.8	Appendix.....	233
15.8.1	Appendix A: Information Flow Overview	233

15.8.2	Appendix B: Prompt Template Development.....	234
15.8.3	Appendix C: Survey Instruments	235
15.8.4	Appendix D: Variable Operationalization.....	237
15.8.5	Appendix E: Coding Schema	238
15.8.6	References	238
16	Publication 7: Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance.....	241
16.1	Introduction.....	241
16.2	Background	243
16.2.1	Individual Creativity Support Systems (ICSS)	243
16.2.2	Generative Language Models (GLMs) and Brainstorming.....	244
16.3	Method	245
16.3.1	System and Usage Process: Brainstorming App with GLM-Capability ..	245
16.3.2	Hypothesis Derivation.....	246
16.3.3	Procedure & Participants.....	248
16.3.4	Treatment & Measurement	248
16.3.5	Analysis.....	250
16.4	Results.....	250
16.4.1	Group Fluency (Number of Ideas)	251
16.4.2	Originality Score	252
16.4.3	Exploratory analysis.....	253
16.5	Discussion	256
16.5.1	General Discussion.....	256
16.5.2	Contribution to Theory.....	259
16.5.3	Implications for Practice	260
16.5.4	Limitations & Further Research.....	260
16.6	Conclusion	262
16.7	References	262
17	Publication 8: Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies.....	267
17.1	Introduction.....	267
17.2	Background	269
17.2.1	Scaffolding.....	269
17.2.2	Generative Language Model (GLM).....	269
17.2.3	DSR Tool Support.....	270
17.3	Method	270
17.4	GLM-based DSR Support Tool	272
17.4.1	Awareness of the Problem	272

17.4.2	Suggestion.....	273
17.4.3	Development	275
17.5	Evaluation	278
17.6	Discussion	282
17.6.1	Meta-level Support.....	282
17.6.2	Content-level Support	282
17.6.3	Limitations	284
17.6.4	Contribution and Outlook	284
17.7	Conclusion	285
17.8	References	285
17.9	Author Biographies	290
17.10	Appendix - Interview Questions	290
18	Publication 9: Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions.....	293
18.1	Introduction	293
18.2	Background	295
18.2.1	Design science research tool support	295
18.2.2	Generative language models	296
18.3	Methodology	297
18.3.1	Study approach.....	298
18.3.2	Survey	300
18.4	Results.....	302
18.5	Discussion	305
18.6	Perception of AI suggestions	305
18.6.1	Usage of AI suggestion in a DSR support tool	306
18.6.2	Towards a DSR support tool with embedded, generative AI capabilities	308
18.7	Limitations	309
18.8	Conclusion and outlook.....	310
18.9	References	311
19	Publication 10: Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy	317
19.1	Introduction	317
19.2	Background	318
19.2.1	Coding (education) in the age of large language models	318
19.2.2	Bloom's (revised) taxonomy	319
19.2.3	Bloom's taxonomy and (coding) learning support tools	320
19.3	Method	320

19.3.1	Data collection	320
19.3.2	Data analysis	321
19.4	Results.....	322
19.5	Discussion	326
19.5.1	General discussion	326
19.5.2	Contributions.....	328
19.5.3	Implications.....	329
19.5.4	Limitations and outlook	329
19.6	Conclusion	331
19.7	References.....	331
20	Publication 11. From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI.....	337
20.1	Introduction.....	337
20.2	Background	340
20.2.1	Foundations of Effort Investment Research.....	340
20.2.2	Effort Investment in Sociotechnical Systems.....	341
20.3	Derivation of the Conceptual Model.....	342
20.4	Method	345
20.5	Results.....	348
20.5.1	Expectancy	349
20.5.2	Instrumentality	351
20.5.3	Valence.....	353
20.5.4	Perceived Responsibility	354
20.5.5	Effort Investment	356
20.5.6	Work Effectiveness and Efficiency.....	357
20.5.7	Cognitive Resources.....	358
20.6	Discussion	359
20.6.1	GenAI's Effect on Human Behavior and Relationships	359
20.6.2	Contribution to Theory.....	364
20.6.3	Implications for Practice	365
20.6.4	Limitations and Further Research	366
20.7	Conclusion	367
20.8	References.....	368
20.9	Appendix.....	371
20.9.1	Appendix A: Interview Guideline	371
20.9.2	Appendix B: Coding Scheme.....	373
	Full publication list	375
	GLM usage while producing this dissertation	377
	Eidesstattliche Versicherung	379

Identität eingereichter Versionen.....	381

I. List of figures

Figure 1. Phenomenon framed as a sociotechnical system in this dissertation.....	9
Figure 2. Integrative overview of the findings across the studies of this dissertation.....	33
Figure 3. Direct and indirect effects of GLMs on task outcomes	35
Figure 4. Disentangled human–GLM dyad performance.....	37
Figure 5. Potential reference points for individual performance evaluation.....	38
Figure 6. Dynamic scaffolding by the GLM.....	40
Figure 7. GLM's role in exploration.....	44
Figure 8. Human creation and curation when working with GLMs	46
Figure 9. Changed effort intensity and direction across job-level tasks	47
Figure 10. Responsibility attribution between the human and the GLM when working on a situated problem-solving task	50

The figures in the publications included in this dissertation are not listed.

II. List of tables

Table 1. Dissertation outline	6
Table 2. Journal publications included in the dissertation	19
Table 3. Conference publications included in the dissertation.....	19
Table 4. First publication: Memmert and Bittner (2022)	21
Table 5. Second publication: Memmert and Tavanapour (2023).....	22
Table 6. Third publication: Memmert, Cvetkovic, and Bittner (2024).....	23
Table 7. Fourth publication: Memmert (2024).....	24
Table 8. Fifth publication: Memmert and Bittner (2024).....	25
Table 9. Sixth publication: Memmert, Cvetkovic, et al. (2025).....	26
Table 10. Seventh publication: Memmert, Mies, and Bittner (2024).....	27
Table 11. Eighth publication: Memmert, Tavanapour, and Bittner (2023)	28
Table 12. Ninth publication: Memmert, Cvetkovic, and Bittner (2023).....	29
Table 13. Tenth publication: Memmert, Borchers, et al. (2025).....	30
Table 14. Eleventh publication: Memmert, Soroko, and Bittner (2025).....	31

The tables in the publications included in this dissertation are not listed.

III. List of abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
DSR.....	Design Science Research
ICSS	Individual Creativity Support System
IS	Information Systems
GLM.....	Generative Language Model
LLM	Large Language Model
ML.....	Machine Learning

The abbreviations in the publications included in this dissertation are not listed.

1 Introduction

1.1 Motivation and problems

Organizations invest heavily in AI (Chakraborty et al., 2024). For structured, repeatable, closed-ended problems—such as image classification or pricing of real estate—AI systems can provide predictions or recommendations that can serve as additional inputs for human judgment and that can lead to significant performance gains (Hemmer et al., 2025), a setup often referred to as AI-assisted decision-making (Bućinca et al., 2021; Dellermann, Ebel, et al., 2019). The rise of generative AI like ChatGPT (T. B. Brown et al., 2020; Feuerriegel et al., 2024) has further expanded AI adoption beyond automation and decision-support into open-ended problem-solving tasks, such as generating ideas, developing concepts, or writing code. By mid-2025, ChatGPT alone was used by roughly 800 million people each week, indicating unprecedented global diffusion and substantial use in both personal and work contexts (Edmonds, 2025). While early evidence points to potential performance improvements across a variety of open-ended tasks (T. B. Brown et al., 2020), emerging research also highlights unintended consequences. Examples include low-quality, AI-generated content (“AI workslop,” Niederhoffer et al., 2025) or more homogeneous output (Dell’Acqua et al., 2023), potentially harming organizational performance (Niederhoffer et al., 2025). Taken together, the rapid and widespread adoption of generative AI, combined with its potential benefits and apparent risks, makes understanding human–GenAI work a practically relevant and timely problem (Challapally et al., 2025).

The adoption of AI for solving problems in practice ties into a broader academic discourse on how humans and AI systems can work together. Scholars have argued that for tasks not suitable for full automation, research should focus on *hybrid intelligence*—that is, human and artificial intelligence working in tandem (Akata et al., 2020; Dellermann, Ebel, et al., 2019), sometimes referred to as human–AI collaboration (Richter & Schwabe, 2025), in which AI acts as a teammate (Seeber et al., 2020). A central expectation for human–AI work includes achieving *superior* performance (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021). While AI has demonstrated high performance in structured, closed-ended decision-making tasks (classification or recommendation), many organizational tasks are open-ended, lack a clear ground truth, and require creativity (Avital & Te’eni, 2009; Berente et al., 2021; Krogh, 2018). Consequently, several scholars have called for exploring how AI can support solving such open-ended problems (Dellermann, Ebel, et al., 2019; Krogh, 2018; Raisch & Fomina, 2025; Seeber et al., 2020).

Alvesson and Sandberg (2011) argue that interesting research challenges traditional assumptions. Following this argument, the above may also be considered an interesting problem from a theoretical perspective, because generative AI may be disruptive by seemingly

enabling new ways in which AI can support humans (content-level contribution) in new kinds of tasks (solving open-ended problems). Traditionally, technology has been regarded as most appropriate for supporting codifiable, repeatable, or routine tasks (Berente et al., 2021; Raisch & Fomina, 2025). Earlier systems could assist open-ended tasks like idea generation at a *meta-level*, for instance, by guiding users through predefined processes such as creativity methods (e.g., Althuizen & Reichel, 2016). At the same time, AI systems could provide *content-level* outputs in narrow, closed-ended tasks (e.g., classification or ranking given a set of predefined options). Generative language models (GLMs), that is, text-based generative AI, however, seem to go further by generating new content—ideas, concepts, code, or texts—for non-routine, open-ended tasks (T. B. Brown et al., 2020). They seem to have what Avital and Te'eni (2009) discussed under the term "generative capacity", that is, an ability to produce something new. Moreover, unlike prior generative systems tailored to specific domains, GLMs, due to their pre-training (T. B. Brown et al., 2020), can be applied across a broad range of tasks without task-specific training (data). Furthermore, they can even generate creative output (Haase & Hanel, 2023), which is essential for problem-solving (Dörner & Funke, 2017), yet traditionally considered a comparative human strength (Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018). This domain- and task-agnostic, (creative) generative capacity challenges what type of organizational tasks can benefit from AI support, or, as Krogh (2018, p. 407) put it, what constitutes a "problematic situation" that AI can meaningfully address.

The "machines as teammates" perspective (Seeber et al., 2020, p. 1) highlights that increasingly capable AI systems may be perceived as partners rather than as mere tools (Rix & Hess, 2022), and, consequently, scholars called for understanding human–AI dynamics (Makarius et al., 2020), and explored how human behavior is shaped by working with AI (e.g., Siemon et al., 2015; Stieglitz et al., 2022). With GLMs dynamically contributing (creative) content akin to another human, the question of how this affects human behavior is relevant (see also Hamza et al., 2024).

In summary, human–GLM problem-solving constitutes a relevant and timely research phenomenon from both a practical and theoretical perspective. In practice, GLMs were adopted broadly (Edmonds, 2025), yet there seems to be a tension between potential performance gains in a variety of open-ended tasks and potential risks (Dell'Acqua et al., 2023; Niederhoffer et al., 2025). From a theoretical perspective, GLMs disrupt how technology can support work, with the capacity to provide (creative) content-level contributions to open-ended problem-solving tasks, which may alter human behavior (Hamza et al., 2024). Consequently, we lack an understanding of how GLMs help or hinder human problem-solving, which mechanisms arise, and under what conditions they unfold.

1.2 Research goal and research questions

Based on the problem discussed above, the following overarching research goal is formulated for this dissertation:

Research Goal. *To explain how humans work with GLMs to solve problems, to uncover the mechanisms that shape performance outcomes, effort, and responsibility in such work, and to understand factors that influence these mechanisms.*

More specifically, the dissertation seeks to answer four interrelated research questions concerning (1) performance outcomes, (2) underlying mechanisms and (3) influencing factors, as well as (4) how these insights may inform designing human–GLM problem-solving.

Working with an AI system is typically oriented toward attaining task-related outcomes (Baird & Maruping, 2021). Scholars have formulated expectations for such ensembles, arguing that humans and AI systems should jointly achieve superior or complementary performance compared to the individual by combining their relative strengths (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021). As discussed above, due to their generative capacity, GLMs appear to be able to support open-ended problem-solving tasks and contribute on a content level in ways that may resemble another human group member, potentially challenging traditional notions of comparative strengths (Dellermann, Ebel, et al., 2019). This introduces uncertainty about how human behavior and joint human–GLM performance are affected when working with such systems, more specifically, whether humans and GLMs jointly outperform solitary humans. Going beyond isolated comparisons of human and AI (e.g., Haase & Hanel, 2023; Koivisto & Grassini, 2023), therefore, the first research question is:

Research Question 1. *How does working with GLMs shape individual and collective performance in problem-solving tasks?*

Understanding performance outcomes is only a first step; it gives rise to the question of why such outcomes occur when humans work with GLMs. Addressing this requires identifying the underlying mechanisms through which GLMs influence problem-solving and human behavior. Prior work has called for research that sheds light on the dynamics of human–AI work (Makarius et al., 2020). For example, recent research has examined whether social mechanisms such as social loafing or responsibility diffusion—commonly observed in human teams—also arise in human–AI collaboration (M. Liu et al., 2023; Siemon et al., 2015; Stieglitz et al., 2022). As discussed above, this appears ever more relevant when working with capable GLMs, which may shape how humans perceive and interact with them (Rix, 2022), potentially inviting parallels to mechanisms known from human group work. Identifying the mechanisms at play when humans engage with GLMs is therefore essential for *explaining* human–GLM problem-solving. This leads to the following research question:

Research Question 2. *What mechanisms arise when humans work with GLMs that shape problem-solving performance and behavior?*

Identifying mechanisms is an important step toward explaining how humans and GLMs jointly solve problems, yet such mechanisms are unlikely to operate uniformly across all situations. Literature suggests that the effects of AI systems on human performance and behavior depend on contextual conditions, including characteristics of the individual, the task, and the interaction setting (Baird & Maruping, 2021; Richter & Schwabe, 2025). In line with this view, several scholars have begun examining who benefits from using GLMs, pointing to heterogeneity across individuals and task contexts (Dell’Acqua et al., 2023). A better understanding of the factors that influence the performance mechanisms discussed above is therefore essential for developing a more comprehensive account of human–GLM problem-solving. This leads to the following research question:

Research Question 3. Which factors influence the performance (mechanisms) that arise when humans work with GLMs?

From a theoretical perspective, explanatory knowledge about consequences and underlying mechanisms is viewed as a foundation for developing design knowledge (Gregor, 2006). Hence, while understanding performance outcomes, mechanisms, and contextual influences is essential for explaining human–GLM collaboration, such insights can also serve as a foundation for informing how systems and practices may be designed to support effective collaboration. Indeed, prior work has emphasized that designing for effective human–AI work constitutes an important research challenge (Akata et al., 2020). This leads to the following research question:

Research question 4. How can systems and practices be designed to foster performance in human–GLM problem-solving tasks?

To maintain a clear focus, the dissertation deliberately delimits its scope along several dimensions. First, following prior scholars who differentiate between problem-solving and decision-making (e.g., Krogh, 2018; also see section 2.1.3), this dissertation focuses on problem-solving, responding to calls to investigate how humans and AI systems jointly address open-ended problems (e.g., Dellermann, Ebel, et al., 2019; Seeber et al., 2020). Second, the dissertation examines human–AI collaboration rather than automation, reflecting arguments that combining human and artificial intelligence represents a central research challenge for solving problems (Dellermann, Ebel, et al., 2019) and is necessary because many organizationally relevant problems are not codifiable (Berente et al., 2021). Third, the primary level of analysis is the human–AI dyad, rather than larger groups. Focusing on dyads is both practically relevant, as many workers use GLMs individually for work, for example, when co-workers are not available (e.g., Dell’Acqua et al., 2023; Retkowsky et al., 2024), and theoretically useful, as it allows for a direct comparison of performance outcomes and helps isolate the behavioral effects of the GLM, avoiding potential behavioral effects due to interactions among humans that may occur in multi-human groups (e.g., Hüffmeier et al., 2013). Fourth, the dissertation focuses on text-based generative AI, specifically GLMs, given their widespread adoption (Edmonds, 2025), the unique ways in which they contribute to open-ended problem-solving (T. B. Brown et al., 2020), and the

fine-grained back-and-forth interaction they allow for (also see section 2.1). Finally, the dissertation emphasizes performance-related outcomes, given their centrality in human–AI collaboration research (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021), while acknowledging that other outcomes like learning or longer-term individual consequences of GLM use, although important, lie beyond its immediate scope.

1.3 Thesis outline

This dissertation is structured as follows. Chapter 2 introduces the necessary terms and concepts and further clarifies the scope of the dissertation that was briefly outlined above. Chapter 3 explains the overarching approach to answering the research questions, relating it to phenomenon-centered research. Chapter 4 provides an overview and briefly describes the eleven first-author publications included in this dissertation. Chapter 5 discusses and interprets the findings across the papers, integrating them to answer the questions posed in the previous section. Chapter 6 highlights the overarching contribution of the dissertation in terms of theory and related discourses in the literature, and in the light of a quickly evolving field, whereas Chapter 7 discusses the practical implications. Chapter 8 discusses the limitations with respect to scope, research design, and theorizing, and points out future areas of research. Chapter 9 draws a brief conclusion. Chapters 0 to 20 contain the full texts of the publications included in this dissertation. The dissertation outline is presented in Table 1.

Table 1. Dissertation outline

Wrapper	1	Introduction
	2	Background
	3	Research design
	4	Publications
	5	Integrative discussion
	6	Theoretical contribution
	7	Practical implications
	8	Limitations and outlook
	9	Conclusion
		References
Publications	10	Paper 1: Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts
	11	Paper 2: Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI
	12	Paper 3: The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3
	13	Paper 4: Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects
	14	Paper 5: Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration
	15	Paper 6: Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load
	16	Paper 7: Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance
	17	Paper 8: Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies
	18	Paper 9: Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions
	19	Paper 10: Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy
	20	Paper 11: From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI

2 Background

2.1 Human–GLM problem-solving

2.1.1 Human–AI collaboration

Terminology. The core phenomenon of this dissertation is human–GLM collaboration for problem-solving tasks, specifically one human working with one GLM to address a specific aspect of a problem. As part of the first publication of this dissertation (Memmert & Bittner, 2022), we identify and define key constructs and the conceptual lens for analysis, which are refined throughout this dissertation. This aligns with the idea of "constant iteration" (Fisher et al., 2021, p. 637) as a key element of phenomenon-centered research adopted in this dissertation.

There are several labels surrounding the idea of human–AI work, including augmentation, collaboration, teaming (or teammates), or hybrid intelligence (Akata et al., 2020; Dellermann, Calma, et al., 2019; Dellermann, Ebel, et al., 2019; Rai et al., 2019; Richter & Schwabe, 2025; Siemon & Wank, 2021). As discussed in Memmert and Bittner (2022), this dissertation builds on Bedwell et al.'s (2012, p. 130) conceptualization of collaboration as an “evolving process whereby two or more social entities actively and reciprocally engage in joint activities aimed at achieving at least one shared goal.” In line with this definition and many other scholars (e.g., Richter & Schwabe, 2025; Zercher et al., 2023), many of the papers in this dissertation frame the respectively investigated human–AI work as collaboration, because functionally the AI system contributes (content) similarly to a human teammate, conditioned on the actions of the human in a back-and-forth manner, towards completing the task. As pointed out by Rix (2022), the perception of a machine as competent or performing well at a specific task increases the likelihood of the system being perceived as a teammate. Likewise, Richter and Schwabe (2025, p. 5) state "while AI may not fulfill the traditional criteria for team membership, it contributes meaningfully to collective outcomes" and propose, although requiring adjustment, "the term 'team' remains relevant in describing human–AI interactions." However, not everyone agrees with conceptualizing AI systems as *teammates*, as discussed by Richter and Schwabe (2025), or working with AI as *human–AI collaboration* (Sarkar, 2023). For example, one might argue that AI systems are not "social entities," or that they cannot "share a goal", yet many studies have shown that humans may respond socially to technical systems (Nass & Moon, 2000), and through prompts, GLMs can be conditioned to produce outputs towards a certain goal (T. B. Brown et al., 2020). Therefore, for this dissertation, we adopt the term 'human–AI collaboration'. Nonetheless, we do not claim that GLMs are equivalent to human teammates; on the contrary, several of our studies report that humans feel differently about an AI as compared to a human teammate.

Performance as an outcome. Several scholars suggested that human–AI collaboration should lead to superior or complementary performance to be *useful* (Vaccaro et al., 2024), that is, the human and AI system should jointly achieve better results compared to the individual entities alone (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021). As part of this dissertation, task performance is therefore prioritized as the outcome, while acknowledging that other outcomes, such as learning, are important as well (Dellermann, Ebel, et al., 2019). More specifically, the dissertation investigates performance when humans and AI systems collaborate, thereby going beyond isolated comparisons of human and AI performance (e.g., Haase & Hanel, 2023; Koivisto & Grassini, 2023).

Dyads as a constellation. A variety of constellations have been discussed in the literature on how humans and AI systems could work together, including individuals or groups of humans with one or multiple AI systems (Schaschek et al., 2024; Zercher et al., 2023). In this dissertation, the focus is set on a group consisting of one human and one AI system (cf. Holzner et al., 2025; Schaschek et al., 2024; Zercher et al., 2023), referred to as a dyad. The dyad was selected for both practical and theoretical reasons. For a practical perspective, while human group work can lead to superior results, other humans are not always cost-effectively available; hence, individuals may turn to AI to (at least) work with the AI instead of alone (Retkowsky et al., 2024). From a theoretical perspective, because we sought to explore the dynamics of humans working with AI systems, as called for in the literature (Makarius et al., 2020). In human groups, social dynamics can affect the group members and their behavior (Hüffmeier et al., 2013; Pinsonneault et al., 1999). Human–AI dyads allow for isolating the cognitive and behavioral mechanisms of human–AI collaboration, avoiding noise associated with human-human social dynamics. Yet, I acknowledge the importance of and emergent discourse on research on one or multiple AI agents being integrated into human groups (Schwabe et al., 2025; Specker et al., 2025; Zercher et al., 2023).

2.1.2 Sociotechnical perspective as a lens

The sociotechnical perspective is well-established in the information systems literature and can serve as a lens for analyzing a social (for example, humans in organizations) and a technical component (for example, an AI-based tool) as a system to achieve a goal or solve a problem (Sarker et al., 2019). The sociotechnical perspective has been used by several scholars for framing and theorizing human–AI work (e.g., Baird & Maruping, 2021; Berente et al., 2021; Dellermann, Ebel, et al., 2019; Maedche et al., 2019; Rix & Hess, 2022; Seeber et al., 2020; Zercher et al., 2023).

As discussed by Sarker et al. (2019), there appears to be no consensus on the specific elements that comprise sociotechnical systems. Following prior research, I conceptualize the phenomenon (human–GLM problem-solving) as a sociotechnical system, in which a human and a technical system (here, GLM) work on a task within a specific context ("situated

task", Baird and Maruping, 2021). Figure 1 presents an illustrative visualization, slightly adapted from an earlier version (Memmert & Bittner, 2022), based on the findings in this dissertation. Task and context were collapsed into "situated task" as discussed by Baird and Maruping (2021, p. 330), as this allows for a clearer visualization of the identified mechanisms, while retaining context as a key influential factor for human–AI work (Baird & Maruping, 2021; Richter & Schwabe, 2025). Moreover, following Baird and Maruping (2021), the outcome dimension was explicitly reflected in the visualization. Due to its central importance for the research phenomenon, performance was selected as the outcome (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021). Performance is defined in relation to the specific tasks in the different studies. The broad arrows indicate (potential) interactions between the elements, whereas the slim arrow indicates the sociotechnical elements leading to the (performance) outcome. Importantly, the goal of this dissertation is not to extend or advance sociotechnical system theory. The sociotechnical perspective is employed as a structuring lens or a vehicle to better understand the underlying phenomenon.

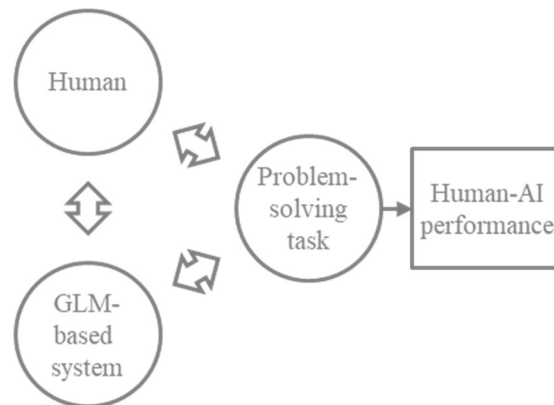


Figure 1. Phenomenon framed as a sociotechnical system in this dissertation

2.1.3 Problem-solving as a type of situated task

While there appears to be no agreed-upon definition of knowledge work, some characteristics include work that is relatively unstructured, situationally contingent, and requires creatively using knowledge to solve problems (Pyöriä, 2005). Pyöriä (2005, p. 123) even argues that "non-routine problem-solving is the core of knowledge work." Problem-solving—the focus of this dissertation—is an essential capability for organizations to innovate (Krogh, 2018) and for individuals to thrive in today's world (Dörner & Funke, 2017). Several scholars called for exploring how problem-solving could be supported by AI (Dellermann, Ebel, et al., 2019; Krogh, 2018; Seeber et al., 2020), even stating "solving complex problems will become the next challenge" of human–AI hybrid intelligence (Dellermann, Ebel, et al., 2019, p. 641).

Several scholars (e.g., Krogh, 2018; Raisch & Fomina, 2025) differentiated between AI for supporting decision-making and supporting problem-solving. It is noteworthy that the problem-solving literature does not necessarily concur with a distinction between problem-

solving and decision-making, and that, for example, decision-making can be considered part of a larger problem-solving process (Quesada et al., 2005). Nonetheless, from a human–AI collaboration perspective, it seems useful to distinguish between clearly defined, repeatable, closed-ended, routine decision-making tasks that are well-suited for classical machine learning-based support (e.g., classification or prediction tasks), and situations that are not well-structured or easily codifiable and traditionally considered difficult to address with technology or AI (also see Berente et al., 2021; Krogh, 2018), or at least, non-generative AI (see next section). Avital and Te'eni (2009) capture a similar line of thought when differentiating between focusing on operational efficiency with convergent thinking toward an outcome whose nature is known in advance, contrasting it with open-ended tasks requiring divergent thinking and creativity toward an outcome that is unknown in nature.

Because it is practically relevant and theoretically interesting, following the above-mentioned calls for research, problem-solving was selected as the focus of this dissertation. However, problem-solving is notoriously difficult to define because it encompasses a broad set of tasks, and different academic traditions have evolved independently (see Dörner & Funke, 2017; Frensch & Funke, 1995; Quesada et al., 2005). Nonetheless, in the context of this dissertation, solving a problem refers to working on a task that requires overcoming certain barriers or uncertainties (Frensch & Funke, 1995), is knowledge-rich (Dörner & Funke, 2017), open-ended or ill-defined (Frensch & Funke, 1995), typically without a ground truth (Dellermann, Ebel, et al., 2019), requires exploring the problem- or solution space and considering the context with multiple interrelated variables, as the path to the solution is not clear up front (Dörner & Funke, 2017; Raisch & Fomina, 2025). While originally focusing on *complex* problem-solving (Memmert & Bittner, 2022), this dissertation uses the term problem-solving as described above to maintain precision, as definitions on what is considered "complex" diverge (Dörner & Funke, 2017). Given that the path towards the solution when solving a problem is typically unknown and needs to be explored, problem-solving seems to require creativity (Dörner & Funke, 2017), for example, generating ideas or options for exploring the problem and solution space (Krogh, 2018; Treffinger et al., 2003).

As part of the broader focus on problem-solving, this dissertation investigates several situated tasks. The focus, with six of the eleven included publications (Memmert, 2024; Memmert & Bittner, 2024; Memmert, Cvetkovic, & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert, Mies, & Bittner, 2024; Memmert & Tavanapour, 2023), is on brainstorming, a specific approach to generating ideas further described in section 2.2.1. As discussed above, generating creative ideas is an important part of solving problems (Treffinger et al., 2003) as potential solutions are typically not known in advance (Raisch & Fomina, 2025). As a second task, this dissertation explores what we referred to as *conceptualizing DSR studies* (Memmert, Cvetkovic, & Bittner, 2023; Memmert, Tavanapour, & Bittner, 2023), referring to developing a (visual) representation of a DSR study, specifying the scope, gath-

ering issues in the literature and in practice, deriving design requirements and design principles, and keeping note of kernel theories; conceptually following the visual approaches by Meth et al. (2015) and Gau et al. (2022). DSR itself was framed as a problem-solving approach before (Thuan & Antunes, 2022), and usually concerns an open-ended problem that requires considering many interrelated elements, thereby reflecting the typical elements of problem-solving as outlined above. A third situated task is developing (including coding) a data science project as part of a university course (Memmert, Borchers, et al., 2025), an open-ended task with many possible solutions and paths towards those solutions, requiring consideration of many interrelated aspects. Lastly, in Memmert, Soroko, and Bittner (2025), professionals describe their use of GLMs for a variety of job-related problems, such as developing strategies, concepts, or presentations; many of which are knowledge-rich tasks that require considering context. Overall, the dissertation covers multiple problem-solving tasks and contexts (i.e., situated tasks).

2.1.4 Generative language models as enabling technology

While scholars have called for exploring how humans can solve problems with the help of AI, this can seem challenging (Berente et al., 2021; Raisch & Fomina, 2025). Technology can be used to support repeatable or recurring tasks, that is, tasks that can be codified, like automating a process (Berente et al., 2021). Moreover, machine learning-based AI systems can be trained on prediction-, ranking-, or classification-problems, that is, for tasks that have a clear input and output, requiring the same decision over and over again, with many past examples of decisions and outcomes, for example, classifying an email as spam. Such tasks, as explained in the prior section, would be considered decision-making tasks. However, organizations are also faced with open-ended problems, and many management or organizational tasks are not codifiable or easily supported by traditional AI systems (Berente et al., 2021; Raisch & Fomina, 2025). In other words, the specifics of the task (problem-solving versus decision-making) pose requirements for the technology to be able to support it. In this dissertation, GLMs are positioned as the key enabler for solving problems with AI, in alignment with recent literature (e.g., Raisch & Fomina, 2025).

GLMs, also referred to as large language models (LLMs), are text-based, machine learning-based systems trained on large corpora of text to predict the next token in a sequence (T. B. Brown et al., 2020). Examples examined throughout this dissertation include GPT-3, GPT-3.5, GPT-4, and ChatGPT. Several of the publications included in this dissertation explicitly position GLMs as a key enabling technology for supporting human problem-solving and outline how GLM-based support differs from that of prior technologies.

In those publications, we argue that the characteristics of problem-solving tasks—being open-ended, ill-defined, knowledge-rich, and lacking a predefined solution space—pose a challenge for traditional technology and discriminative AI, which Banh and Strobel (2023, p. 63) define as AI that "process[es] data to conduct classification, regression, or cluster and determine decision boundaries," and which is well-suited for codifiable, repeatable or

routine, or narrow decision tasks (Banh & Strobel, 2023; Krogh, 2018; Raisch & Fomina, 2025). In contrast, GLMs enable direct support for such tasks because they can generate novel, task-specific content without requiring task-specific training or a predefined ground truth (e.g., Memmert, Cvetkovic, & Bittner, 2023; Memmert & Tavanapour, 2023). Instead, they can be dynamically conditioned on user input at inference time, that is, after the pre-training, during usage, enabling adaptive human–AI interaction (Memmert, Cvetkovic, et al., 2025). This enables GLMs to engage *reciprocally* with the human, which is an important facet of collaboration (Bedwell et al., 2012).

Building on this observation, several papers in this dissertation introduce and elaborate on the notion of "content-level support" (e.g., Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023; Memmert, Tavanapour, & Bittner, 2023) to distinguish GLM-based support from prior technological support. Content-level support refers to the GLM's ability to contribute content, like ideas or alternatives, that are conditioned on the specific problem instance (i.e., a specific brainstorming question), rather than merely structuring or facilitating the problem-solving process at a meta-level, that is, independently of a specific brainstorming question (Althuizen & Reichel, 2016). This distinction is used throughout the dissertation to explain why GLMs can participate in tasks such as brainstorming, conceptualizing DSR, or coding in a new way.

The dissertation and many included papers deliberately use the term GLM instead of LLM because it emphasizes the core enabling capacity, its generativity. Moreover, it uses GLM instead of generative AI, because generative AI is broader, also encompassing other modalities such as image or video generation, which are outside the scope of this dissertation.

Throughout the papers, we argue that this new way of support is why GLMs can be considered a "qualitative leap" (Memmert, Cvetkovic, et al., 2025). Because GLMs can contribute content to a task and adapt to the specific situation and the human in a back-and-forth, reciprocal manner, they may functionally resemble a human collaborator, which in turn may influence the dynamics of human–AI work (see section 2.3). Understanding these dynamics is therefore the focus of this dissertation.

2.2 Creative idea generation

2.2.1 Brainstorming

Creativity is important for problem-solving and innovation in organizations (Amabile, 1988), for example, when exploring different options or solution paths (Dörner & Funke, 2017). Creativity refers to the ability to produce ideas that are original (or surprising or novel) and useful (or effective or valuable) (Amabile, 1988; Runco & Jaeger, 2012).

Brainstorming is a popular method for generating creative ideas (Maaravi et al., 2021) and was hence used as a foundational approach for six of the eleven studies included in this

dissertation. Brainstorming was suggested as a method for human groups to generate creative ideas, following specific rules such as encouraging wild ideas and focusing on quantity of ideas (Osborn, 1953). Brainstorming supports what can be described as divergent thinking, that is, generating many different ideas (Guilford et al., 1961), as opposed to convergent thinking, like evaluating and selecting ideas. Established metrics to assess brainstorming performance are, for example, fluency (number of ideas) or flexibility (e.g., number of different categories of ideas covered) for the pool of generated ideas within a brainstorming session, and novelty or value for the individual ideas themselves (Dean et al., 2006; Dugosh et al., 2000; Nijstad et al., 2010). Over many decades, scholars have developed an in-depth understanding of the mechanisms in human group brainstorming (Karau & Williams, 1993; Pinsonneault et al., 1999) and have developed technological support for both human groups and individuals (K. Wang & Nickerson, 2017).

2.2.2 Individual creativity support systems

Decades of (information systems) research have explored how technology can aid individuals in creative tasks like generating ideas. These systems can be referred to as individual creativity support systems (ICSS, K. Wang & Nickerson, 2017). Traditionally, creativity is framed as a comparative human strength (Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018). Hence, ICSS may focus on supporting the ideation process, for example, guiding users through a creativity method (meta-level support), or, at most, offering stimuli during ideation sessions, while leaving the generation of substantive content largely to the human user (Althuizen & Reichel, 2016; K. Wang & Nickerson, 2017, 2019).

In several of the papers included in this dissertation, we argue that GLM-based ICSS are different from prior (what we refer to as) meta-level ICSS or stimuli-provider ICSS (Memmert, Cvetkovic, et al., 2025). More specifically, we argue that they can offer content-level support by dynamically generating complete, creative ideas in response to the current state of the brainstorming session (Memmert, Cvetkovic, et al., 2025), making them an interesting object of study with respect to creative performance and user behavior.

Prior literature pointed out that individual differences influence creativity (Amabile, 1988). The question of who benefits from ICSS support, for example, the less or the more creative, has since been investigated; however, the results are inconclusive (K. Wang & Nickerson, 2017). Given that GLMs contribute substantive content rather than merely structuring the creative process, understanding how individual differences shape users' interactions with and benefits from such systems is important.

2.3 Mechanisms in (human) groups

Decades of research have explored various mechanisms and effects that occur when humans work together as part of a group. These mechanisms are of particular interest because

they influence individual behavior and performance, as well as the overall group performance (Pinsonneault et al., 1999). In the case of group brainstorming, examples of those mechanisms (sometimes referred to as process gains or process losses; see Pinsonneault et al., 1999) include, cognitive stimulation, that is, a performance-enhancing mechanism in which ideas contributed by one group member stimulate further ideation by others, leading to more or more creative ideas (Dugosh et al., 2000; Pinsonneault et al., 1999). Another example relevant to brainstorming, as well as to group work more generally, is social loafing, which refers to a reduction in effort when working as part of a group (Karau & Williams, 1993; Latané et al., 1979; Pinsonneault et al., 1999). A potential antecedent for such social loafing is diffused responsibility, that is, feeling less responsible for the outcome when working as a group; another antecedent is reduced expectancy or instrumentality, referring to feeling that the own effort is less relevant for achieving the outcome (Karau & Williams, 1993; Pinsonneault et al., 1999).

Given that humans are known to potentially socially respond to technical systems (Nass & Moon, 2000), and, as explained above, machines or AI systems might be perceived as teammates, particularly when becoming more capable (Benz et al., 2024; Rix & Hess, 2022; Seeber et al., 2020), scholars have explored whether similar mechanisms arise when a human works with an AI system instead of another human. For example, researchers demonstrated that ICSS can also induce the positive effects of cognitive stimulation, enhancing ideation performance (Althuizen & Reichel, 2016). Others explored whether and when negative mechanisms like social loafing occur when working with AI (Rix & Hess, 2022; Siemon et al., 2015; Stieglitz et al., 2022), and with GLMs specifically (Hamza et al., 2024). Stieglitz et al. (2022) even proposed a new, adjusted loafing concept—smart loafing—defined as "the reduction of effort in human-[virtual assistant] collaboration to maintain cognitive resources and enhance efficiency in work" (Stieglitz et al., 2022, p. 758), suggesting that relying on an AI system may be desirable when performance is not negatively affected.

Given that GLMs dynamically contribute content functionally similar to another human, for example, ideas at a creative level comparable to many human ideas (Haase & Hanel, 2023), we use these well-documented mechanisms from human groups as a starting point for analyzing human–GLM work. Importantly, this dissertation does not assume that these mechanisms directly transfer to human–GLM dyads, but uses them to structure the empirical investigation, seeking to examine whether, how, and under which conditions they materialize in or require adjustment for human–GLM problem-solving.

3 Research Design

3.1 Phenomenon-centered research

This dissertation adopts an empirical, phenomenon-centered approach to investigating how humans work with GLMs to solve problems, because it aligns well with the research goal set out for the dissertation, and because it was suggested by scholars as an approach for theorizing emerging technologies, including AI (Gregory & Henfridsson, 2021; Krogh, 2018).

Phenomenon-centered research is not a fixed process or method, but a stance in which the phenomenon leads the theorizing, often through challenging existing assumptions and offering new explanations for an undertheorized phenomenon. What constitutes a phenomenon is not clearly defined. Some scholars use a broad conceptualization; for example, Monteiro et al. (2022) use the terms emerging technologies and digital phenomena interchangeably, and Krogh (2018, p. 404) refers to "artificial intelligence" as a "rapidly emerging phenomenon [...]". Others, like Fisher et al. (2021, p. 632), have a more restrictive understanding, defining phenomena as "regularities that are unexpected, that challenge existing knowledge (including the extant theory) and that are relevant to scientific discourse." For this dissertation, I deliberately adopted the broader, former understanding of what constitutes a phenomenon, allowing for the exploration of more specific facets throughout the dissertation.

Phenomenon-centered research is well-suited for theorizing digital phenomena, including AI (Gregory & Henfridsson, 2021; Krogh, 2018). It seems particularly fitting for exploring GLMs and problem-solving, because, as discussed before, GLMs have capabilities like generativity (Avital & Te'eni, 2009) that allow them to support new kinds of tasks (open-ended problems) in new ways (content-level), which may change how these technologies shape performance and human behavior, which is why I adopted this approach for this dissertation.

Phenomenon-centered research focuses on developing a deep understanding of new phenomena, defining what makes them different from prior phenomena, and exploring the key elements, mechanisms, and boundaries thereof (Fisher et al., 2021). This aligns well with the research interest of the dissertation, namely, the goal of gaining an in-depth understanding of human–GLM problem-solving and the four research questions: understanding human–GLM performance, the underlying mechanisms, and boundary conditions, and how these insights can inform design.

While the phenomenon-centered approach does not prescribe a fixed process, several recommendations or sensibilities developed by scholars (Fisher et al., 2021; Monteiro et al.,

2022) are implemented in this dissertation, including connecting to prior theory (section 3.2) and engaging in an empirical dialog and being methodologically flexible (section 3.3).

3.2 Sociotechnical framing of the phenomenon

"Connecting the phenomenon with exciting theories and perspectives" is an important aspect of the phenomenon-centered research (Fisher et al., 2021, p. 633). As pointed out by Monteiro et al. (2022, p. ii), using the phenomenon-centered approach when analyzing digital phenomena requires "fine-grained sociotechnical insight." As explained in the section 2.1.2, this dissertation, following prior research (Baird & Maruping, 2021; Berente et al., 2021; Dellermann, Ebel, et al., 2019; Maedche et al., 2019; Rix & Hess, 2022; Seeber et al., 2020), frames the phenomenon as a sociotechnical system, in which a human and a technical system (here, GLM) work on a task within a specific context ("situated task", Baird & Maruping, 2021, p. 330) to achieve a certain outcome (here, performance).

When analyzing phenomena, Fisher et al. (2021, p. 637) recommend to "have a focal theory and a complementary theory that together explain the phenomenon." As explained above, the sociotechnical perspective guides our conceptualization and analysis of the phenomenon. For more specific facets of the phenomenon, for example, to understand how working with the GLM may influence human behavior, more specific theories, for example, on social loafing, cognitive stimulation, or cognitive load, are employed. Following Fisher et al. (2021), these act as complementary lenses to explain particular mechanisms.

3.3 Empirical, multi-method approach

To get a deep understanding of the phenomenon, Monteiro et al. (2022, p. ii) suggested to "engage in a dialogue with the empirical setting," while being "methodologically responsive." Human-GLM problem-solving is a multi-faceted phenomenon with a sociotechnical nature. Moreover, GLMs and their adoption for work are an emerging and quickly evolving field, which poses additional requirements for the approach.

Therefore, to address the research goal, this dissertation employs a multi-method approach across multiple contexts, communicating the results in multiple papers. Conducting empirical research is important to ground the phenomenon in real-world observation (Monteiro et al., 2022). Moreover, Monteiro et al. (2022, p. ii) suggest being "methodologically responsive." Similarly, Venkatesh et al. (2013) and Mingers (2001) highlight the potential of mixed-method research, combining qualitative and quantitative modes of inquiry for advancing theory. Hence, ten of the papers included in this dissertation report on empirical studies, in which we used a variety of methods, from qualitative research methods to gain an in-depth understanding of experiences when working with GLMs, to quantitative experimental studies to understand and measure performance implications.

Phenomena-centered research is closely aligned with the research questions of this dissertation. A key aspect of phenomena-centered research is to identify a new phenomenon, understand its key elements, and explain how it differs from prior concepts (Fisher et al., 2021; Monteiro et al., 2022). Research question 1 focuses on understanding how working with GLMs shapes individual and collective performance in problem-solving tasks. It seeks to clearly define how working with GLMs is different from prior technology, and particularly, how performance is affected. Due to its importance for human–GLM problem-solving (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021; Krogh, 2018), performance was discussed in all included publications.

Central to gaining an in-depth understanding of a phenomenon is to uncover the underlying mechanisms that drive it and their relevant boundaries (Fisher et al., 2021). This aligns well with the idea of mechanism-based explanations of Hedström and Ylikoski (2010, p. 50), who suggest that it is important to understand the "cogs and wheels" of a system, as well as with the idea of a "theory for explaining and predicting" (Type IV theory, Gregor, 2006, p. 633). Research question 2, which seeks to uncover the mechanisms that arise when humans work with GLMs and that shape problem-solving performance and behavior, and research question 3, focusing on the relevant factors affecting performance, directly align with those core aspects of phenomenon-centered research.

Lastly, developing a deep understanding of a phenomenon and the underlying mechanisms can be useful to develop design knowledge (Gregor, 2006). Thus, research question 4 is concerned with how systems and practices can be designed to foster performance in human–GLM problem-solving tasks, responding to several calls on developing knowledge on designing human–AI work (Akata et al., 2020; Seeber et al., 2020).

Following the idea of phenomenon-centered research and to answer these research questions, the dissertation employs a combination of methods. This includes a literature review to analyze the state of human–AI problem-solving and identify potential research contexts, or, in the words of Baird and Maruping (2021, p. 330), "situated tasks" suitable for empirically investigating human–AI problem-solving (Memmert & Bittner, 2022). For two studies, we systematically generated content to explore the feasibility of using GLMs to support brainstorming (Memmert, Cvetkovic, & Bittner, 2024) or DSR concept development (Memmert, Cvetkovic, & Bittner, 2023). Several studies follow using purpose-built GLM-based web applications (Memmert, 2024; Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert, Mies, & Bittner, 2024; Memmert & Tavanapour, 2023) or commercial, off-the-shelf GLMs (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025) for collecting data from participants on them working with GLMs. This included responses to qualitative questionnaires after lab experiments or as part of a semester-long study on coding with GLMs (Memmert, Borchers, et al., 2025), or in-depth interview studies on GLM use by knowledge workers in industry (Memmert, Soroko, & Bittner, 2025). Moreover, several studies used quantitative data, including both self-reports and log data from the interaction with the GLMs. This includes qualitative and quantitative evaluations

that can inform design decisions. Participants were predominantly students and knowledge workers in professional settings.

Using a variety of methods and data sources not only aligns with the call to be "methodologically responsive" (Monteiro et al., 2022, p. ii), but also with good scientific practice referred to as data and method triangulation (Döring & Bortz, 2016), which offers a more comprehensive perspective on the phenomenon and can increase the robustness of the results. As described in section 2.1.3, we explored different situated tasks, which aligns with scholars' suggestion to use different contexts to enhance external validity (Gregory & Henfridsson, 2021), that is, to explore how the phenomenon unfolds in different contexts. Moreover, using different methods and data sources can help to counter weaknesses, such as subjectivity of self-reported data. Moreover, it can lead to uncovering tensions and new, interesting phenomena, for example, when subjective perception and measured performance do not align.

To summarize, this dissertation follows a phenomenon-centered research approach, seeking to gain an in-depth explanatory understanding of the human–GLM problem-solving phenomenon, using a multi-method, multi-study empirical approach. Because of the quickly evolving nature of the field, a cumulative form of a dissertation was chosen to iteratively communicate the results and take part in the emerging scientific discourse.

4 Publications

4.1 Overview

To address the research goal, as explained in the prior chapter, this dissertation follows a multi-study, multi-paper approach. This dissertation contains eleven publications, of which three are journal articles (Table 2), and eight are conference articles (Table 3, seven completed papers, one short paper). All papers were published at reputable and peer-reviewed journals and conferences including *Business & Information System Engineering* (BISE, 2 articles), *Journal of Information System Education* (JISE, 1 article), *International Conference on Information Systems* (ICIS, 1 completed paper), *European Conference on Information Systems* (ECIS, 2 completed paper, 1 short paper), and *Hawaii Conference on Information Systems* (HICSS, 4 completed paper). Lucas Memmert is the first author on all eleven included publications; the respective co-author contributions are described in the following section.

Table 2. Journal publications included in the dissertation

No.	Publication	Section
6	Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load	15
8	Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies	17
11	Effort Management of Professionals when Working with Generative Language Models	20

Table 3. Conference publications included in the dissertation

No.	Publication	Section
1	Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts	10
2	Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI	11
3	The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3	12
4	Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects	13
5	Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration	14
7	Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance	16
9	Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions	18
10	Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy	19

4.2 Key information on the publications

The following section contains a short overview of each of the papers included in the dissertation, including meta information like the ranking of the outlet, the type of the paper,

track, and mini-track (if applicable), co-author contribution statement, as well as the abstract and a short explanation how it relates to the other publications in this dissertation and the dissertation as a whole. The individual contributions per publication are listed according to the author contribution taxonomy (NISO CRediT Working Group). All co-authors have signed to confirm that these statements are accurate and that they agree with them. All articles are reproduced verbatim, including appendices, at the end of this document. To ensure a consistent appearance, they have been reformatted.

4.2.1 Paper 1: Memmert and Bittner (2022)

Table 4. First publication: Memmert and Bittner (2022)

Full Citation	Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In <i>Proceedings of the 55th Hawaii International Conference on System Sciences</i> , Virtual Event/Maui, HI, USA. http://hdl.handle.net/10125/79376
Abstract	Solving complex problems has been proclaimed as one major challenge for hybrid teams of humans and artificial intelligence (AI) systems. Human-AI collaboration brings immense opportunities in these complex tasks, in which humans struggle, but full automation is also impossible. Understanding and designing human-AI collaboration for complex problem solving is a wicked and multifaceted research problem itself. We contribute to this emergent field by reviewing to what extent existing research on instantiated human-AI collaboration already addresses this challenge. After clarifying the two key concepts (complex problem solving and human-AI collaboration), we perform a systematic literature review. We extract research contexts and assess them considering different complexity features. We thereby provide an overview of existing and guidance for designing new, suitable research contexts for studying complex problem solving through human-AI collaboration and present an outlook for further work on this research challenge.
In-Text Citation	(Memmert & Bittner, 2022)
Link/DOI	http://hdl.handle.net/10125/79376/
VHB-Rating 2024	B
Type	Conference Paper Completed Research Paper
Track	Collaboration Systems and Technologies Collaboration with Intelligent Systems: Machines as Teammates
Role in this dissertation	This literature review (22 publications included) clarifies the phenomenon and defines relevant terms, such as human-AI collaboration and (complex) problem-solving. It identifies relevant "situated tasks" (i.e., research contexts) and potential technical enablers for human-AI problem-solving (i.e., generative AI), offering early insights into the human-AI problem-solving landscape, and inspiring the subsequent empirical studies (and selected situated tasks).
Co-authors' contribution	Lucas Memmert: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. Eva Bittner: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

4.2.2 Paper 2: Memmert and Tavanapour (2023)

Table 5. Second publication: Memmert and Tavanapour (2023)

Citation	Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In <i>Proceedings of the 31st European Conference on Information Systems (ECIS)</i> , Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
Abstract	Groups of humans or crowds can be remarkable when coming up with ideas. However, not everyone has a group of humans at their disposal to brainstorm. With recent advances in AI, however, generative large language models (LLM) might be capable of contributing ideas in a brainstorming session, turning individual work of a human into joint work of human and AI. It is, however, unclear, how group effects known from human brainstorming groups transfer to such a human-AI setting. In our mixed-method study (qualitative emphasis) with 24 participants, we investigate how a human brainstorms together with the generative LLM ‘GPT-3’, and how they perceived their experience. Our results highlight known effects like cognitive stimulation but also a risk of free riding. We thereby contribute to the understanding of how generative AI, which is becoming broadly available, can be used to address the challenge of human-AI collaboration for solving open-ended problems.
In-Text Citation	(Memmert & Tavanapour, 2023)
Link/DOI	https://aisel.aisnet.org/ecis2023_rp/429
VHB-Rating 2024	A
Type	Conference Paper Completed Research Paper
Track	Human-AI collaboration: Challenges and opportunities
Role in this dissertation	This mixed-methods study (n=24) argues that GLM-based creativity support is fundamentally different (content-level support) and explores the perceptions and usage behavior of individuals when working with a GLM-based, purpose-built brainstorming prototype, following our prior identification of “creative content generation” as a potentially interesting research context in Memmert and Bittner (2022). It explores and reflects on mechanisms known from human groups when brainstorming with GLMs, highlighting both positive effects, such as cognitive stimulation, but also potentially negative side effects, including reduced effort (akin to social loafing) or responsibility diffusion. Moreover, it empirically shows the potential performance improvements through AI (GLM) in creative tasks.
Co-authors’ contribution	Lucas Memmert: Conceptualization (lead), Data curation, Formal analysis, Investigation, Methodology (lead), Project administration, Software, Visualization, Writing – original draft, Writing – review & editing (lead). Navid Tavanapour: Conceptualization (support), Methodology (support), Resources, Supervision, Writing – review & editing (support)

4.2.3 Paper 3: Memmert, Cvetkovic, and Bittner (2024)

Table 6. Third publication: Memmert, Cvetkovic, and Bittner (2024)

Citation	Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In <i>Proceedings of the 57th Hawaii International Conference on System Sciences</i> , Honolulu, HI, USA. https://hdl.handle.net/10125/107289
Abstract	Generative language models (GLM) like GPT-3 can support humans in creative tasks. Such systems are capable of generating free-text output based on a provided input prompt. Given the outputs' sensitivity to the prompt, many techniques for prompt engineering were proposed both anecdotally in social media and increasingly in literature. It is, however, unclear if and how such a system and such techniques can be employed in creative contexts such as for generating ideas. In our study, we investigate the effects of using six prompt engineering techniques. For each combination of techniques, we have GPT-3 generate ideas for an exemplary scenario. The ideas are rated according to novelty and value. We report on the effects of the (combinations of) prompt engineering techniques. With our study, we contribute to the emerging field of prompt engineering and shed light on supporting idea generation with GLMs, showing a pathway to embedded GLM capabilities.
In-Text Citation	(Memmert, Cvetkovic, & Bittner, 2024)
Link/DOI	https://hdl.handle.net/10125/107289
VHB-Rating 2024	B
Type	Conference Paper Completed Research Paper
Track	Software Technology Generative and Conversational AI in Information Systems Research and Education: Opportunities and Challenges
Role in this dissertation	A key design decision for GLM-based ICSS is the embedded prompt template. This study conducts a systematic, quantitative evaluation of 64 prompt templates, contributing to the early discussion on prompt engineering. The findings inform the design of the prompt templates for the purpose-built brainstorming prototypes used in several of the brainstorming studies in this dissertation.
Co-authors' contribution	Lucas Memmert: Conceptualization, Data curation, Formal analysis (support), Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft (lead), Writing – review & editing (lead). Izabel Cvetkovic: Conceptualization, Data curation, Formal analysis (lead), Investigation, Methodology, Writing – original draft (support), Writing – review & editing (support). Eva Bittner: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

4.2.4 Paper 4: Memmert (2024)

Table 7. Fourth publication: Memmert (2024)

Citation	Memmert, L. (2024). Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects. In <i>Proceedings of the 32nd European Conference on Information (ECIS)</i> , Paphos, Cyprus. https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1
Abstract	Generative language models (GLMs) like GPT-3.5 or ChatGPT produce remarkable results when used for creative tasks. In this short paper, we outline our planned approach to investigating how humans brainstorm together with GLMs. Particularly, we plan to explore how the presentation of GLM suggestions affects brainstorming group effects such as cognitive stimulation or social loafing known from all-human groups. Based on group brainstorming literature and previous studies, we designed an experiment to measure underlying performance effects. In the planned between-subject experiment, participants brainstorm with a GLM. We investigate how manipulating the completeness and the communicated origin of suggestions affects performance-related constructs. In this short paper, we only report preliminary results. Our study has theoretical implications explaining how well-documented brainstorming group effects from all-human groups can help to understand and affect human-AI team performance. Practical implications include insights into designing GLM-based creativity support while balancing performance dimensions.
In-Text Citation	(Memmert, 2024)
Link/DOI	https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1
VHB-Rating 2024	A
Type	Conference Paper Short Paper
Track	Human-AI Collaboration
Role in this dissertation	A key concern for GLM-based support is ensuring appropriate engagement with and reliance on GLM suggestions. This study develops a conceptual model for an intervention and reports on results from a prior experiment (n=26), partially covering the conceptual model for humans brainstorming with GLMs.
Co-authors' contribution	(Single-author paper)

4.2.5 Paper 5: Memmert and Bittner (2024)

Table 8. Fifth publication: Memmert and Bittner (2024)

Citation	Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In <i>Proceedings of the 57th Hawaii International Conference on System Sciences</i> , Honolulu, HI, USA. https://hdl.handle.net/10125/106425
Abstract	With the widespread adoption of generative large language models (GLMs) such as GPT-3 or ChatGPT for human-AI problem solving, understanding the effect on performance becomes important. Brainstorming is an established approach for generating ideas to solve problems. In this study, we investigate how AI ideas affect the brainstorming performance metric ‘flexibility’, which refers to the breadth of exploration or coverage of the topic. The foundation for our analysis is the data from an experiment (n=52) in which individual participants brainstormed in two conditions: (1) human-only (baseline) and (2) human+AI (treatment). The treatment condition had access to ideas generated via the GLM OpenAI GPT-3.5. Results show significantly higher flexibility for the human+AI as compared to the human-only condition with a large effect size. With our study, we contribute to the literature of electronic brainstorming, brainstorming with GLMs, as well as to the research challenge of human-AI collaboration.
In-Text Citation	(Memmert & Bittner, 2024)
Link/DOI	https://hdl.handle.net/10125/106425
VHB-Rating 2024	B
Type	Conference Paper Completed Research Paper
Track	Collaboration Systems and Technologies Collaboration with Intelligent Systems: Machines as Teammates
Role in this dissertation	Our previous study (Memmert & Tavanapour, 2023) and prior literature render the question of whether GLM support leads to a broader or narrower exploration of a brainstorming question uncertain. In this between-subject experiment (n=52) using a purpose-built brainstorming prototype, we empirically demonstrated that human–GLM dyads outperform solitary humans in terms of flexibility (i.e., diversity of ideas) and showed that GLMs drive category exploration.
Co-authors’ contribution	Lucas Memmert: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing. Eva Bittner: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

4.2.6 Paper 6: Memmert, Cvetkovic, et al. (2025)

Table 9. Sixth publication: Memmert, Cvetkovic, et al. (2025)

Citation	Memmert, L., Cvetkovic, I., Tavanapour, N., & Bittner, E. (2025). Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load. <i>Business & Information Systems Engineering</i> . Advance online publication. https://doi.org/10.1007/s12599-025-00974-y
Abstract	The adoption of generative AI (artificial intelligence) has sparked increased interest in how individual humans can generate ideas and solve problems collaboratively with AI, aiming to achieve superior performance compared to working alone. Based on group effects from all-human brainstorming groups, the paper investigates a human–AI brainstorming setting using OpenAI’s GPT-3.5 as an embedded generative large language model (GLM). In a between–subjects experiment (n = 75) comparing solitary humans and human–GLM dyads, results show that humans do not perform better individually when supported by a GLM. However, collectively, the human–AI dyad achieves superior (or complementary) performance on common brainstorming performance metrics (fluency, flexibility, novelty, and value). Findings are discussed in relation to effort allocation and potential loafing behavior when working with GLMs. The paper advances our understanding human–AI group dynamics, the transferability of group mechanisms (cognitive stimulation, free riding, cognitive inertia) from all-human to human–AI groups, and the discourse on smart loafing with AI group members.
In-Text Citation	(Memmert, Cvetkovic, et al., 2025)
Link/DOI	https://doi.org/10.1007/s12599-025-00974-y
VHB-Rating 2024	B
Type	Journal Paper
Track	-
Role in dissertation	This between-subject experimental study (n=75) using a purpose-built brainstorming prototype is a quantitative exploration and confirmation of the qualitative findings of Memmert and Tavanapour (2023). It expands the study of Memmert and Bittner (2024) through additional metrics, data, and theorizing to gain a detailed understanding of human-GLM brainstorming in terms of brainstorming performance (fluency, flexibility, novelty, and usefulness) and cognitive load. It empirically demonstrates the superior performance of the human-GLM dyads compared to solitary brainstormers, explores effects on human behavior, and sheds light on the transferability of human group mechanisms in human-GLM dyads.
Co-authors’ contribution	Lucas Memmert: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. Izabel Cvetkovic: Data curation. Navid Tavanapour: Data curation, Investigation. Eva Bittner: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

4.2.7 Paper 7: Memmert, Mies, and Bittner (2024)

Table 10. Seventh publication: Memmert, Mies, and Bittner (2024)

Citation	Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In <i>Proceedings of the 45th International Conference on Information Systems (ICIS)</i> , Bangkok, Thailand. https://aisel.aisnet.org/icis2024/aibus/aibus/7/
Abstract	Generative language models (GLMs) are used for creative tasks such as brainstorming. While studies show GLMs' creative potential, research on who benefits from support in creative tasks is inconclusive, and concerns are raised on whether humans reduce their effort when working with GLMs. Therefore, we explored how creative ability and GLM support affect the two brainstorming performance metrics fluency and originality. In our experiment with a GPT-4-based brainstorming app, we randomly assigned the 58 participants to one of two conditions: with or without GLM-support. Additionally, we split participants into a low- and a high-creativity group based on a creativity ability pre-test. Our findings reveal that low-creativity participants show increased fluency when being GLM-supported, while the effect on originality was positive for both groups. We contribute to the literature on human-GLM brainstorming and individual creativity support systems, offering a more nuanced perspective on GLMs' (potentially undesired) effects on behavior and performance.
In-Text Citation	(Memmert, Mies, & Bittner, 2024)
Link/DOI	https://aisel.aisnet.org/icis2024/aibus/aibus/7/
VHB-Rating 2024	A
Type	Completed Research Paper
Track	AI in Business and Society
Role in this dissertation	This between-subject experimental study (n=58) using a purpose-built brainstorming prototype offers a quantitative examination of individual differences, specifically, individual creativity, following up on prior literature and own qualitative findings on the role of task-related ability when using or benefitting from GLMs (Memmert, Cvetkovic, & Bittner, 2023; Memmert & Tavanapour, 2023; Memmert, Tavanapour, & Bittner, 2023).
Co-authors' contribution	Lucas Memmert: Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing. Julian Mies: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing – original draft. Eva Bittner: Funding acquisition, Writing – review & editing

4.2.8 Paper 8: Memmert, Tavanapour, and Bittner (2023)

Table 11. Eights publication: Memmert, Tavanapour, and Bittner (2023)

Citation	Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. <i>Journal of Information Systems Education</i> , 34(3), 279–292. https://aisel.aisnet.org/jise/vol34/iss3/3
Abstract	Design science research (DSR) is taught in university courses or used by students for their final theses. For successfully learning DSR, it is important to learn to apply it to (real-world) problems. However, students not only need to learn the new DSR paradigm (meta-level) but also need to develop an understanding of the problem domain (content-level). In this paper, we focus on content-level support (CLS), proposing an illustrative tool to aid students when learning to develop a conceptual design with DSR (e.g., for a prototype). Following the DSR paradigm, we deductively identify students' issues and use the scaffolding approach to develop design requirements (DRs) and design principles (DPs). To offer AI-generated scaffolding, we use the generative language model (GLM) 'GPT-3'. We evaluate our illustrative design through 13 expert interviews. Our results show that providing students with CLS is perceived to be helpful, but the interaction with the student needs to be designed carefully to circumvent unintended usage patterns. We contribute DPs and an illustrative instantiation thereof toward a DSR tool support ecosystem. More broadly, we contribute to the understanding of how humans can be supported by AI to solve problems, an important challenge in human-AI collaboration research.
In-Text Citation	(Memmert, Tavanapour, & Bittner, 2023)
Link/DOI	https://aisel.aisnet.org/jise/vol34/iss3/3
VHB-Rating 2024	C (Note: AIS Special Interest Group Education (SIG ED) Recommended Journal)
Type	Journal Paper
Track	---
Role in this dissertation	This early study, inspired by the potential of AI to support design (Memmert & Bittner, 2022), follows the DSR paradigm to develop and evaluate (interviews, n=13) design principles for a tool for supporting students in developing visual conceptualizations for their DSR study (e.g., issues, design requirements, design principles) with the support of GLMs. The study sought to explore how GLMs could support problem-solving and is the conceptual counterpart to the technical feasibility study by Memmert, Cvetkovic, and Bittner (2023). It inspired subsequent studies on cognitive stimulation to extend problem-space exploration through GLMs (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023)
Co-authors' contribution	Lucas Memmert: Conceptualization (lead), Data curation, Formal analysis, Investigation, Methodology (lead), Project administration, Software, Visualization, Writing – original draft, Writing – review & editing (lead). Navid Tavanapour: Conceptualization (support), Methodology (support), Supervision, Writing – review & editing (support). Eva Bittner: Conceptualization (support), Funding acquisition, Methodology (support), Supervision, Writing – review & editing (support)

4.2.9 Paper 9: Memmert, Cvetkovic, and Bittner (2023)

Table 12. Ninth publication: Memmert, Cvetkovic, and Bittner (2023)

Citation	Memmert, L., Cvetkovic, I., & Bittner, E. (2023). Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions. In <i>Proceedings of the 31st European Conference on Information Systems (ECIS)</i> , Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/405
Abstract	Solving complex problems was named a new challenge for research on human-AI collaboration. In our study we focus on a particular means to solving complex problems: design science research (DSR). We investigate whether AI, more specifically, generative language models (GLM), can support an individual in conceptualizing DSR studies by making helpful suggestions. To do so we use extracts of a published DSR study and have GPT-3, a GLM provide suggestions for aspects of this study. These suggestions are then evaluated in a survey (n=33) regarding their helpfulness. Results show that GLM suggestions are perceived to be helpful, with some variation depending on expertise. Reported interest in using such a tool in the future was high. Describing how GLMs can offer helpful suggestions we contribute toward a DSR tool support ecosystem and, more generally, towards knowledge on how humans and (generative) AI systems can team up to solve complex problems.
In-Text Citation	(Memmert, Cvetkovic, & Bittner, 2023)
Link/DOI	https://aisel.aisnet.org/ecis2023_rp/405
VHB-Rating 2024	A
Type	Completed Research Paper
Track	Design Research and Methods in Information Systems
Role in this dissertation	This early study investigates whether GLM can generate helpful content for developing concepts for DSR studies, serving as the technical counterpart to the conceptual study by Memmert, Tavanapour, and Bittner (2023), who developed a tool design. It systematically generates DSR concept elements such as issues, design requirements, and design principles using GPT-3 and analyzes and reports on the helpfulness of these generated elements, based on data collected via a quantitative survey (n=33).
Co-authors' contribution	Lucas Memmert: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft (lead), Writing – review & editing (lead). Izabel Cvetkovic: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft (support), Writing – review & editing (support). Eva: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

4.2.10 Paper 10: Memmert, Borchers, et al. (2025)

Table 13. Tenth publication: Memmert, Borchers, et al. (2025)

Citation	Memmert, L., Borchers, M., Plückhahn, J., & Bittner, E. (2025). Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy. In <i>Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)</i> , Waikoloa Village, HI, USA. https://hdl.handle.net/10125/109714
Abstract	With the advancement of large language models (LLMs) such as ChatGPT, students increasingly employ LLMs during university courses to improve their coding skills. However, it is still unclear what using these systems means for student learning. In our study, we explore how students employ a LLM when solving a realistic coding challenge over the course of a semester in small groups. We analyze students' LLM chat log data through the framework of Bloom's taxonomy, a well-established education framework, to understand their problem-solving and learning behavior. We enrich our analysis with students' responses to weekly survey questions, contextualizing our findings with subjective experiences. We find that students primarily use LLMs for lower-order thinking skills like remembering, understanding, and applying procedural knowledge. With our study, we contribute to the growing literature on understanding the effects of working and learning with LLMs and offer practical suggestions for teachers and students.
In-Text Citation	(Memmert, Borchers, et al., 2025)
Link/DOI	https://doi.org/10.24251/HICSS.2025.862
VHB-Rating 2024	B
Type	Completed Research Paper
Track	Software Technology Generative and Conversational AI in Information Systems Research and Education: Opportunities and Challenges
Role in this dissertation	Following our earlier study on supporting students with GLMs (Memmert, Tavanapour, & Bittner, 2023), this study investigates how students utilize GLMs when working on semester-long coding projects. It employs a diary study (n=31) with weekly qualitative surveys and examines the chat logs, finding various ways in which GLMs scaffold students' problem-solving, similar to our earlier study on DSR support (Memmert, Tavanapour, & Bittner, 2023) and professionals using GLMs (Memmert, Soroko, & Bittner, 2025).
Co-authors' contribution	Lucas Memmert: Conceptualization, Data curation, Formal analysis (lead), Methodology (lead), Project administration, Software, Supervision, Validation, Visualization, Writing – original draft (lead), Writing – review & editing. Marten Borchers: Conceptualization, Funding acquisition, Investigation, Methodology (support), Writing – original draft (support). Juliane Plückhahn: Data curation, Formal analysis (support), Writing – original draft (support). Eva Bittner: Conceptualization, Funding acquisition

4.2.11 Paper 11: Memmert, Soroko, and Bittner (2025)

Table 14. Eleventh publication: Memmert, Soroko, and Bittner (2025)

Citation	Memmert, L., Soroko, D., & Bittner, E. (2025). From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI. <i>Business & Information Systems Engineering</i> (67), 615–635. https://doi.org/10.1007/s12599-025-00960-4
Abstract	Generative Artificial Intelligence (GenAI) is adopted by knowledge workers to boost productivity, yet its specific characteristics such as probabilistic outputs and human-level content generation may change how professionals think about their effort. Prior literature has warned about unintended side effects of AI, but experiments on effort reduction when working with AI – which could threaten performance – reported mixed results. GenAI's rapid adoption combined with its specific characteristics make it critical and timely to clarify how GenAI influences knowledge workers' effort in professional settings. The qualitative study draws on 21 interviews with knowledge workers who frequently use GenAI for work. A directed content analysis, guided by expectancy theory and social loafing frameworks, revealed that most interviewees do not simply reduce effort, but rather strategically reallocate or even increase effort. They continuously learn to steer GenAI, viewing themselves as process administrators. The traditional group-based mechanisms of reduced effort or diffused responsibility do not seem to be directly transferable to human–GenAI dyads in professional settings. By revealing that GenAI reshapes the factors that influence effort rather than simply eroding motivation, providing a multifaceted view of effort investment beyond mere reduction, and highlighting the interplay between human relationships and GenAI-facilitated work, this research advances the discourse on human-(Gen)AI dynamics and the unintended consequences of (Gen)AI. Recognizing these shifts when setting policies and expectations enables organizations to benefit from GenAI's potential while mitigating potential risks to performance.
In-Text Citation	(Memmert, Soroko, & Bittner, 2025)
Link/DOI	10.1007/s12599-025-00960-4
VHB-Rating 2024	B
Type	Journal Research Paper
Role in this dissertation	Following several prior studies (Memmert, Tavanapour, & Bittner, 2023), this qualitative interview study (n=21) conceptualizes effort and its antecedents and consequences in relation to GLM use among knowledge workers in professional contexts. Thereby, the study, for example, broadens the focus from loafing behavior (see, e.g., Memmert, Cvetkovic, et al., 2025) to effort management and uncovers boundary conditions, such as for responsibility attribution (see, e.g., Memmert & Tavanapour, 2023).
Co-authors' contribution	Lucas Memmert: Conceptualization (lead), Data curation, Formal analysis (lead), Investigation, Methodology, Project administration, Visualization, Writing – original draft (lead), Writing – review & editing (lead). Daria Soroko: Conceptualization (support), Data curation, Formal analysis (support), Investigation, Methodology, Writing – original draft (support), Writing – review & editing (support). Eva Bittner: Conceptualization (support), Funding acquisition, Supervision, Writing – review & editing (support).

5 Integrative discussion

The sociotechnical perspective was introduced as a lens to conceptualize the phenomenon and is subsequently used to answer the four research questions. The sections below offer the overarching, integrated discussion on the human–GLM problem-solving phenomenon across the eleven studies (see Figure 2). Section 5.1 covers how GLMs shape performance outcomes. Section 5.2 then theorizes underlying mechanisms linking these elements of the sociotechnical system. Section 5.3 discusses factors influencing the performance and the identified mechanisms. Section 5.4 discusses specific design knowledge and implications derived from the findings in this dissertation.

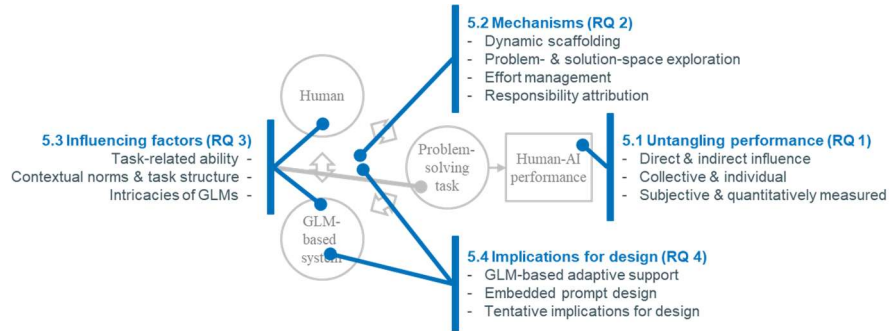


Figure 2. Integrative overview of the findings across the studies of this dissertation

5.1 Untangling human–GLM task performance (RQ 1)

A core research challenge for human–AI collaboration is for the human and AI to achieve superior or complementary performance as compared to the individual when solving problems (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021; Vaccaro et al., 2024). The situated tasks investigated as part of this dissertation concern open-ended problems, requiring creativity, which is traditionally considered a comparative strength of humans (Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018). The first research question pertains to how GLMs shape problem-solving performance.

5.1.1 Facets of performance

Because achieving superior or complementary performance is such a central ambition of human–AI collaboration, this research question was considered in all papers included in the dissertation, combining different perspectives and methods. Collectively, these studies examined both the expected *potential* of GLMs for enhancing problem-solving outcomes and the *realized* effects when humans actually worked with them.

- *Anticipated helpfulness.* We designed an illustrative GLM-based tool and asked experts for their perspectives on how such a system could particularly benefit novices (Memmert, Tavanapour, & Bittner, 2023).

- *Technical feasibility.* In two experiments without direct human tool interaction, we examined the capability of GLMs to generate text that human raters considered creative or helpful within specific problem-solving contexts, thereby demonstrating the potential of GLMs for content-level support (Memmert, Cvetkovic, & Bittner, 2023, 2024).
- *Theorized improvement pathways.* Building on prior literature, a conceptual model was developed that articulates how GLMs might influence performance (Memmert, 2024).
- *Subjective performance perceptions.* We asked participants about their experiences when brainstorming with our self-developed GLM-based prototype or working with commercial, off-the-shelf GLMs. These qualitative studies captured perceived performance impacts across creative and professional task contexts (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025; Memmert & Tavapour, 2023).
- *Quantified performance differences.* Finally, we conducted controlled experiments comparing individuals and human–GLM dyads, showing quantitatively that GLMs can improve task performance, although a more nuanced analysis is required to understand how these effects emerge (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). Moreover, we conducted a controlled, between-subject experiment to understand the effect of individual differences (Memmert, Mies, & Bittner, 2024).

Together, these studies provide a multi-angle perspective on how working with GLMs shapes problem-solving performance. The following sections synthesize the findings.

5.1.2 GLM's direct and indirect performance influence

The studies show that GLMs contribute directly and indirectly to task outcomes, as depicted in Figure 3. Prior systems, such as in ICSS, typically facilitated human problem-solving at a meta-level, for example, by guiding the user through a pre-defined process (Althuizen & Reichel, 2016). By contrast, GLMs can provide content-level contributions that are themselves perceived as creative or helpful (Memmert, Cvetkovic, & Bittner, 2023, 2024). In other words, GLMs can directly generate task-relevant material that becomes part of the work product (see Figure 3)—such as ideas, text segments, or code snippets. For example, in the case of idea generation, GLMs can produce ideas of comparable creative quality to many humans (Haase & Hanel, 2023; Holzner et al., 2025) and contribute them directly to the joint pool of ideas. This ability to contribute creative output directly, rather than merely facilitate a process, challenges long-held assumptions that creativity is an exclusively human strength (cf. Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018).

working with GLMs produced significantly fewer ideas than solitary participants (Memmert, Cvetkovic, et al., 2025). This finding is contrary to expectations from traditional ICSS, where exposure to external stimuli typically improves idea generation (Althuizen & Reichel, 2016), meaning that it is expected that ICSS improve the performance of the human. Unlike our GLM-based ICSS, these earlier ICSS typically do not provide full ideas themselves (e.g., Althuizen & Reichel, 2016). Our findings seem to confirm concerns of reduced individual performance raised in the literature (e.g., M. Liu et al., 2023).

If humans contribute fewer ideas when working with a GLM, this can have negative consequences. Studies show that the most creative ideas are still contributed by humans (Haase & Hanel, 2023; Koivisto & Grassini, 2023), and that while GLM ideas might seem promising at first sight, the assessment may change when executing on those ideas (Si et al., 2025). Furthermore, humans may possess unique or situational knowledge inaccessible to the GLM (see also Fügener et al., 2021). Therefore, when humans become passive and contribute fewer ideas, relying on the GLM to produce ideas, this may adversely affect results.

Comparing the performance of humans working without interaction with others—referred to as a *nominal* group—to humans working with others—referred to as a *real* group—is typical in brainstorming research to investigate group mechanisms (Pinsonneault et al., 1999). However, applying this evaluation logic without adjustment to human–AI dyads may lead to mischaracterizing the situation (Memmert, Soroko, & Bittner, 2025), as discussed in detail in section 5.2.3. Our findings underscore the importance of disentangling individual and collective performance to understand how GLMs shape collaborative outcomes (Figure 4). They also motivate a deeper exploration of the underlying cognitive and motivational mechanisms, discussed in the section 5.2. It is important to note that conceptualizing performance differently for other task types might be more appropriate (Karau & Williams, 1993).

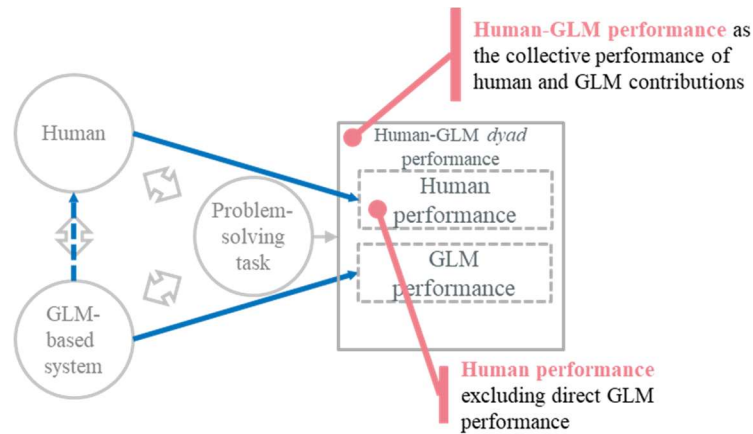


Figure 4. Disentangled human–GLM dyad performance

5.1.4 Subjective vs. quantitatively measured performance

In several studies, participants reported experiencing performance improvements when collaborating with the GLM (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023). These subjective perceptions, however, did not always align with quantitatively measured performance outcomes. Despite some participants explicitly stating that the GLM helped them think of additional ideas or new directions, individuals working with GLMs did not outperform (or even performed worse on fluency than) solitary humans on an individual level when GLM-generated ideas were excluded (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025), that is, when considering what was described as individual performance in the section above. This may be interpreted as a *perception–performance gap*, where perceived improvement does not fully correspond to measured outcomes. The studies provide several complementary interpretations for this gap.

Participants often described brainstorming or problem-solving as “easier” or “faster” when the GLM was available (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023), consistent with prior literature (e.g., Specker et al., 2025). This subjective sense of increased efficiency may have contributed to more positive subjective performance evaluations, even in the absence of measurable individual gains. This would be consistent with prior literature on human groups that showed that humans brainstorming in groups perceived their performance as better compared to nominal brainstormers, despite not actually performing better on an individual level (e.g., P. B. Paulus et al., 1993).

Moreover, participants working with the GLM may have used the joint output as a reference point for evaluation (Figure 5). While participants working with the GLM did not generate all ideas themselves, they requested, evaluated, and selected ideas produced by the GLM. In other words, rather than being *purely creators*—as in the solitary condition—they acted as both *creators and curators*. Participants integrated GLM-generated content into their final outputs (Memmert, Cvetkovic, et al., 2025). As a result, they may have evaluated their performance based on the jointly produced outcome (Memmert, Cvetkovic,

et al., 2025). This hybrid creation–curation role became even clearer in the study with professionals, where participants iteratively refined GLM outputs alongside their own contributions (Memmert, Soroko, & Bittner, 2025). Evaluating performance at the level of the combined output could explain why participants felt effective, even when their individual performance decreased in some metrics (Memmert, Cvetkovic, et al., 2025). This interpretation also aligns with the process perspective: for all studies, the GLM's input entering the final work result was contingent upon the human actively requesting and integrating it and for the professional use and coding studies (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025), participants were even required to formulate one or multiple prompts. Therefore, overall performance was indeed *contingent on their usage behavior*. Hence, even when not attributing the GLM contributions to themselves directly, humans did contribute to the work result beyond the ideas they created by integrating GLM output into the work result.

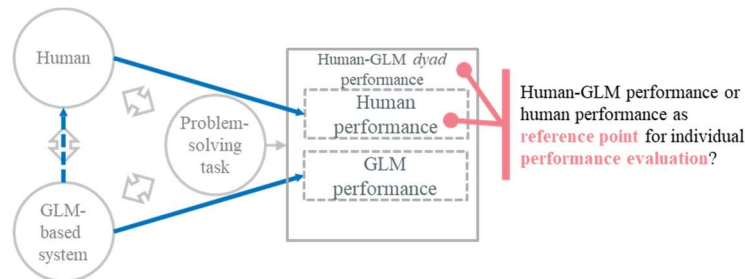


Figure 5. Potential reference points for individual performance evaluation

Besides the perspective of who contributed what, there is also the question of how others, particularly the recipients, perceive and evaluate work results. In a professional context (Memmert, Soroko, & Bittner, 2025), participants clearly stated that managers, colleagues, and clients ascribe the performance to them, regardless of whether or not they have used a GLM. Hence, according to others, there is no differentiation between what the human contributed and what the GLM contributed, potentially leading participants to evaluate their performance similarly, that is, at the level of the joint outcome.

5.1.5 Summary of performance implications

Overall, results seem to suggest that GLM use influences (collective) task performance positively. We empirically showed that human–GLM dyads can achieve superior performance in problem-solving compared to the individual, both subjectively from the worker perspective and when measuring (collective) performance, i.e., GLMs have—at least in the short term and from the perspective of the individual—positive effects with respect to task performance. This is consistent with the literature (e.g., Dell’Acqua et al., 2023; Holzner et al., 2025). However, GLM use may have negative effects for the individual, or beyond the dyad, for both instrumental and humanistic goals, like (potentially) reduced understanding and learning (Memmert, Borchers, et al., 2025; Memmert, Tavanapour, & Bittner, 2023), negative long-term individual effects (Kosmyna et al., 2025), or homogenization of results (see also Dell’Acqua et al., 2023; Holzner et al., 2025, see section 5.2.2).

Untangling performance, our studies jointly surfaced three key findings. First, **GLMs can contribute both directly** by adding content to the outcome, **and indirectly**, by influencing the human, their contribution, and behavior. The specific characteristics of GLMs enable this support in areas like creativity, which were traditionally considered a human strength (Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018). Second, **differentiation between individual and collective human–GLM performance** allows for unpacking the specific effects of GLMs on work. While collectively GLM–human dyads seem to outperform solitary humans, the humans themselves do not perform better or even worse when working with a GLM compared to unaided humans, requiring further inquiry. Third, closely related to the prior point, there appears to be an interesting tension between the **subjectively perceived and quantitatively measured performance** effect of GLMs. While participants may have felt like they performed better, their individual performance may have been unchanged, or they may have actually even contributed less.

5.2 Mechanisms: Effects on behavior and problem-solving performance (RQ 2)

The second research question pertains to the discourse on how GLMs shape human behavior and outcomes, specifically seeking to understand the emerging dynamics or underlying mechanisms in human–AI dyads (see Makarius et al., 2020). The prior section already offered some more specific insights into the performance components themselves. Yet, merely showing that GLMs and humans can achieve superior performance only draws a partial picture, because how the sociotechnical system functions largely remains a black box. Therefore, the following section details four mechanisms identified across the studies, offering tentative insights into the interaction between the human, the GLM-based system, and the task within a context (situated task), on how these interactions lead to the observed outcomes, such as superior performance. The mechanisms are: (1) dynamic scaffolding, (2) problem-and solution-space exploration, (3) effort management, and (4) responsibility attribution. While not fully explored in this dissertation, mechanisms are important foundations for designing artifacts and hence lay a crucial foundation for future, design-oriented work (Gregor, 2006).

5.2.1 Dynamic scaffolding

Scaffolding is a technique to aid particularly novices and learners to perform tasks or solve problems they otherwise might not be able to perform or solve (Quintana et al., 2004; Reiser, 2004). Scaffolding may be static, e.g., a checklist of steps to complete, or dynamic, e.g., a teacher or tutor providing situation-specific cues or hints (Saye & Brush, 2002). We argue that its generative ability allows GLMs to provide dynamic scaffolding in problem-solving situations, that is, to provide adaptive, context-specific support that helps humans initiate and progress through problem-solving tasks (Memmert, Tavanapour, & Bittner, 2023). This dynamic support differs from traditional, predefined scaffolds (e.g., checklists)

because GLM responses are context-aware and generated dynamically in real-time. While discussed in detail in Memmert, Tavanapour, and Bittner (2023), the dynamic scaffolding potential of GLMs surfaced throughout multiple papers, with the GLM repeatedly fulfilling different scaffolding functions, including overcoming the blank page problem (e.g., Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023), providing psychological safety (Memmert, Cvetkovic, et al., 2025), or supporting operationalization (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025). In terms of the sociotechnical system perspective, the GLM influences the relationship between the human and the task through its dynamic scaffolding (see Figure 6).

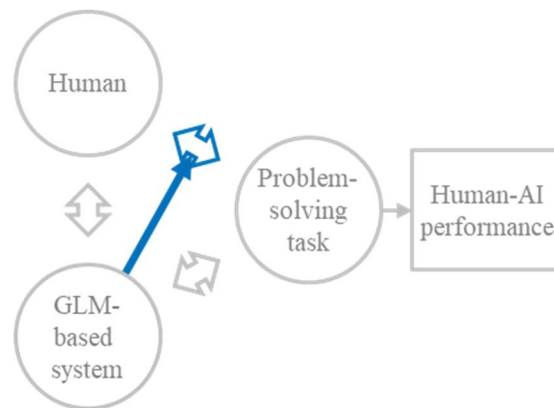


Figure 6. Dynamic scaffolding by the GLM

The **blank-page problem** refers to the challenge of getting started on a task when starting with an empty page (Oliver, 1982). Across several papers, participants reported feeling that the GLM offered initial content, for example, initial ideas during a brainstorming session (Memmert & Tavanapour, 2023) or initial code when starting to develop software (Memmert, Borchers, et al., 2025). This initial basis allowed them to then come up with their own content, or adapt or improve upon the GLM suggestion (Memmert, Borchers, et al., 2025), i.e., get started quickly instead of wasting time before putting something on paper (Memmert, Tavanapour, & Bittner, 2023). Our studies suggest that this might be especially helpful for novices (Memmert, Borchers, et al., 2025; Memmert, Tavanapour, & Bittner, 2023). Dynamic scaffolding is not limited to the beginning of the task. Solving problems may require considering a variety of factors and being knowledgeable in different areas (Frensch & Funke, 1995; Krogh, 2018). For example, in the case of conducting a DSR study, one inter alia needs to be knowledgeable about the domain, the DSR paradigm, and potentially relevant theories from adjacent fields. This can be particularly difficult for novices (Memmert, Tavanapour, & Bittner, 2023). In those situations, GLMs may support by providing suggestions which then serve as a starting point for own research or ideas. The GLM can thereby prevent the human from getting stuck and instead moving efficiently towards the goal, fulfilling what can be referred to as "direction maintenance" scaffolding function (Wood et al., 1976, p. 98). Yet, as pointed out by our interviewees, such usage is associated with certain risks (e.g., hallucinations), which is why GLM outputs should be treated as

hypotheses requiring validation, specifically when aiding research activities (Memmert, Tavanapour, & Bittner, 2023).

Beyond mere content-level scaffolding, participants report that the GLMs provided a feeling of **(psychological) safety** when working on a task. With the GLMs providing initial input or examples (Memmert, Borchers, et al., 2025; Memmert, Cvetkovic, et al., 2025; Memmert, Tavanapour, & Bittner, 2023), it reduced hesitation and uncertainty among participants, fulfilling a "demonstration" scaffolding function (Wood et al., 1976, p. 98). For example, in a brainstorming study, some participants reported being unsure about the intended direction for the brainstorming session and therefore using the GLM to orient themselves. However, there are caveats to the psychological safety dynamic scaffolding facet. At least in our studies, the GLM did not receive more information about the task than the GLM (both human and GLM received the same task description). Thus, while for some participants receiving GLM suggestions seems to provide psychological safety, it is not assured that the GLM ideas indeed point in a useful direction. Hence, the use of GLMs could lead to a false sense of security.

GLMs can also provide more concrete support when working on a task. A recurring facet is what might be referred to as **operationalization support**, or what Kim and Hannafin (2011, p. 407) refer to as "taking over lower-order tasks." The GLM helps participants to translate abstract or conceptual ideas into concrete artifacts. This facet of the dynamic scaffolding mechanism, in the context of software development, materialized in students knowing in principle what they wanted to achieve with their software, but lacking the specific knowledge to express it in code (Memmert, Borchers, et al., 2025). The GLM effectively helped to bridge the gap between conceptual design and implementation. A similar operationalization support was also reported by participants when performing knowledge work (Memmert, Soroko, & Bittner, 2025), for example, letting GLMs handle lower-level formulation so the human focuses on higher-level concept development.

While such scaffolding might be a pathway to performance complementarity, with the human and GLM focusing on their respective strengths (cf. Dellermann, Ebel, et al., 2019), the degree of scaffolding varies with how the GLM is prompted. Embedded designs, in which the interaction is abstracted via a graphical user interface, can offer a way to combine static and dynamic scaffolding (e.g., Memmert, Tavanapour, & Bittner, 2023). In contrast, free-text interaction, like with ChatGPT, requires the user to effectively prompt the GLM. For example, in our study (Memmert, Borchers, et al., 2025), participants rarely used the GLM for having their code reviewed, even though GLMs can offer good feedback (Popovici, 2024). Hence, dynamic scaffolding is contingent on the interaction, more specifically, on the user's awareness of the GLM's capabilities and the user's ability to use them through appropriate prompting.

Overall, dynamic scaffolding explains how GLMs provide responsive guidance throughout the problem-solving process. In our study (Memmert, Tavanapour, & Bittner, 2023), we

highlight the potential for using GLMs in research and for teaching or supervision in learning environments. Since then, several studies have further discussed or explored GLM for supporting research (Dwivedi et al., 2023; Si et al., 2025; Susarla et al., 2023) and leaning (Jiang & Nakatani, 2025; Sun & Deng, 2025), or discussed the specific scaffolding paths, like overcoming the blank page problem by "acting as an icebreaker" (Schwabe et al., 2025, p. 369). The dynamic scaffolding mechanism interacts with the effort management mechanism described in section 5.2.3.

5.2.2 Problem- and solution space exploration

The problem- and solution-space exploration mechanism refers to how GLM contributions, and interacting with GLM ideas, changes how broadly, deeply, or creatively problem-solving unfolds. Literature suggests that GLMs, due to their generative ability, can support humans in more broadly exploring both the problem and the solution space (Bouschery et al., 2023), i.e., they can support the divergent part of the problem-solving process. Already in our early literature review, we showed that the potential of the generative ability of generative AI was explored for the creative content creation domain (Memmert & Bittner, 2022). In follow-up studies, we show that experts, in principle, see potential for GLMs to support problem solvers (Memmert, Tavanapour, & Bittner, 2023), and that GLMs can indeed provide content considered creative or helpful (Memmert, Cvetkovic, & Bittner, 2023, 2024), also see section 5.1.

An advantage of group brainstorming compared to individual brainstorming is that the ideas of one group member can inspire another group member to think of additional ideas or think in new directions (Pinsonneault et al., 1999) by triggering associative thinking (Dugosh et al., 2000), thereby expanding search in the problem or solution space. This phenomenon is known as cognitive stimulation, and is considered a classical group brainstorming process or process gain (Coskun et al., 2000; Dugosh et al., 2000; Pinsonneault et al., 1999). Cognitive stimulation can lead to enhanced coverage of the brainstorming question, i.e., a broader exploration (Dugosh et al., 2000). ICSS were developed to produce a similar effect, namely, to enhance the brainstorming performance of individuals by providing external stimuli (Althuizen & Reichel, 2016; K. Wang & Nickerson, 2017). Given that our earlier studies demonstrated that GLMs can produce content that is considered creative, we explored whether a similar mechanism to cognitive stimulation occurs when humans work with GLMs.

Across multiple studies, participants reported to have felt inspired by the GLM ideas and that they helped them to discover new perspectives, consistent with prior literature (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023; Schwabe et al., 2025; Specker et al., 2025). This suggests that the classical human-group brainstorming mechanism of cognitive stimulation, which is present in stimuli-provider ICSS (Althuizen & Reichel, 2016), may also occur when the brainstorming counterpart is

a GLM instead of a human. However, we would then expect that the individual brainstorming performance improved, i.e., that humans supported by GLMs produce more, more novel, more useful, or more diverse ideas (Dugosh et al., 2000). But when measuring individual performance (e.g., fluency, cf. Elshan et al., 2025), humans themselves did not produce more creative ideas when working in groups compared to when working alone. This is contrary to what would be expected if cognitive stimulation were to occur. Indeed, several participants reported feeling that working with the GLM limited their creativity, anchoring them to certain ideas (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). This seems to resemble what is referred to as cognitive inertia or fixation, that is, exposure to external stimuli, particularly when closely related (Baruah & Paulus, 2016), can limit exploration (see Jansson & Smith, 1991; Nijstad & Stroebe, 2006; K. Wang & Nickerson, 2019). Given that creativity was traditionally considered a human strength (Dellermann, Ebel, et al., 2019), GLMs reinforcing existing ideas may not have been surprising; nonetheless, it seems contrary to the reports of participants feeling inspired.

As we observed that the collective human–GLM dyad's creative performance was higher compared to the solitary brainstormer's performance (see section 5.1), we explored who drove this performance improvement as part of a follow-up analysis. We find that GLM ideas were rated more novel on average (Memmert, Cvetkovic, et al., 2025; consistent with, e.g., Haase & Hanel, 2023; Koivisto & Grassini, 2023), and that it was mostly the GLM—not the human—that suggested ideas in new categories, i.e., that drove the problem-space exploration (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). This makes AI-induced cognitive inertia seem less likely. However, these findings are interesting in light of studies that have found working with GLMs can lead to the homogenization of results (Dell'Acqua et al., 2023; Doshi & Hauser, 2024; Holzner et al., 2025). While *individuals* benefit from GLM support, comparing pooled outputs of individuals and pooled outputs of human–GLM pairs may show that the outputs in the latter pool are more alike (Dell'Acqua et al., 2023; Doshi & Hauser, 2024).

Given these results, it appears that cognitive stimulation did not unfold in a similar manner when working with GLMs compared to when working in human groups or when using classical stimuli-provider ICSS. Future research should explore whether and when cognitive stimulation occurs, and whether it is counteracted by other mechanisms, such as distraction or changed effort (see section 5.2.3). Moreover, future research should explore appropriate interaction design to realize the benefits of GLM stimulation while minimizing potential negative side effects, for example, following Buçinca et al.'s (2021) work on increasing engagement with AI, as also discussed by Memmert (2024).

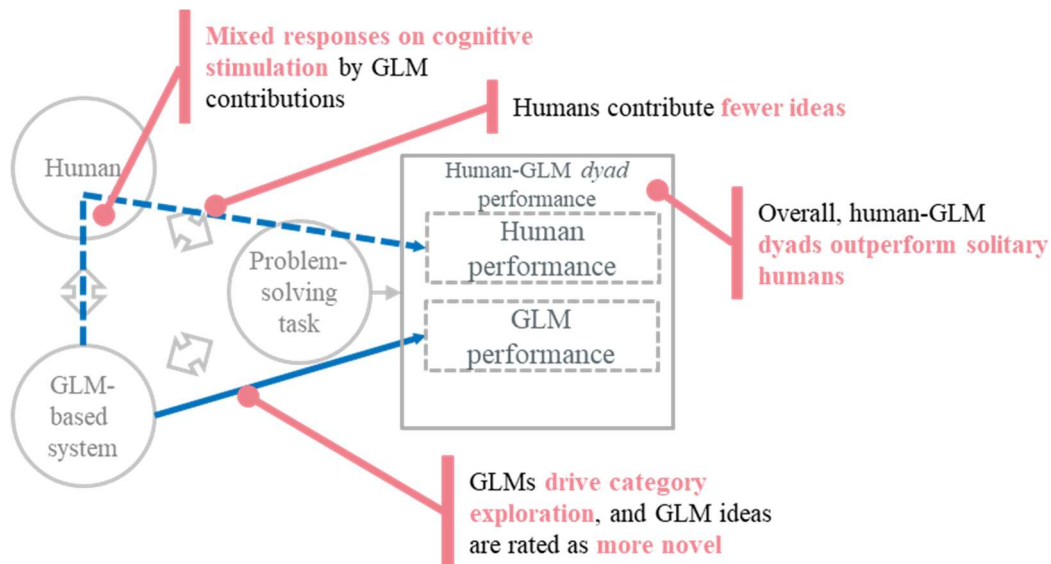


Figure 7. GLM's role in exploration

Importantly, and somewhat surprisingly, given that traditionally creativity is considered a human strength (Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018), GLMs act as drivers of broader exploration in problem and solution spaces (as suggested by Bouschery et al., 2023); section 5.4 offers tentative design implications. However, while diversity might be increased on an individual level, when pooling ideas, GLMs may lead to homogenization (Doshi & Hauser, 2024). This could potentially have implications for organizations' competitive advantage (Holzner et al., 2025); if everyone receives similar GLM output, it may prevent the exploration of options that set companies apart.

5.2.3 Effort management

Human effort when working with GLMs. A concern in the literature on human–AI collaboration is that working with AI systems may have undesired side effects, such as reduced human motivation or effort (Siemon et al., 2015; Stieglitz et al., 2022). Drawing on literature on social loafing, which refers to reduced individual effort in group settings (Latané et al., 1979), scholars have suggested that social loafing-like behavior may occur when humans collaborate with AI (M. Liu et al., 2023; Siemon et al., 2015) and GLMs specifically (Hamza et al., 2024), and that humans may not invest effort but uncritically accept AI suggestions (Krogh, 2018). Such effects could be problematic from a hybrid intelligence perspective, as performance complementarity relies on the combination of human and AI strengths rather than substituting human effort (Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021).

Several findings in our studies are consistent with these concerns. When confronted with a specific GLM-based support tool, experts anticipated that such tools could lead (particularly novice) users to accept AI suggestions without sufficient reflection (Memmert, Tavapour, & Bittner, 2023). Empirically, in the brainstorming studies, a substantial share of GLM-generated suggestions was accepted without modification (Memmert, Cvetkovic,

et al., 2025; Memmert, Mies, & Bittner, 2024). Participants' subjective reports further indicate perceived reductions in effort, with one participant explicitly stating that they had "become lazy" when working with the system (Memmert, Cvetkovic, et al., 2025). These perceptions are reflected in a quantitative metric, fluency (number of ideas), commonly used to assess effort when brainstorming (Elshan et al., 2025; Nijstad et al., 2010; P. Paulus, 2000), humans working with a GLM produced fewer ideas (Memmert, Cvetkovic, et al., 2025). At face value, this may resemble social loafing-like behavior, manifesting in a reduction in direct human contributions. Qualitative evidence provides additional insight into the motivational dynamics underlying this effect. Participants described how classical antecedents of social loafing, such as *expectancy* and *instrumentality* (Karau & Williams, 1993), were weakened in the presence of GLMs (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025). For example, some reported that GLM-generated ideas appeared superior to their own or that their contribution felt redundant (reduced expectancy), and others perceived the GLM output as already sufficient for many situations, diminishing the perceived necessity of (additional) own input (reduced instrumentality). Taken together, these observations suggest that GLMs can reduce the exerted observable human effort in problem-solving tasks.

Effort management across levels. While these observations seemingly align with social loafing, the behavior may require a more nuanced interpretation. Rather than a uniform reduction in effort, workers appear to deliberately and situationally redistribute effort across sub-tasks, their job, and time horizons, which we describe as effort management (Memmert, Soroko, & Bittner, 2025). This interpretation builds on Stieglitz et al. (2022, p. 758) newly introduced "smart loafing" concept for human-AI collaboration, suggesting that relying on the AI may be desirable if performance is not harmed and cognitive resources can be maintained. We conceptualize effort management across three interrelated levels.

First, at the **intra-task level**, humans shift from being the sole creator to being both a creator and a curator (Figure 8). While most participants continued to contribute at least some content themselves, many spent time and effort reviewing, validating, or adjusting the GLM output (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025; also see section 5.2.4 on responsibility). Hence, on the task level, the seeming reduction in effort might occur because the GLM performs part of the sub-tasks (e.g., idea generation, code or concept drafting), whereas the human may focus on other parts, like the big picture or reviewing and integrating content (also see Retkowsky et al., 2024 on reconfigured roles). This aligns with the concept of performance complementarity and with the dynamic scaffolding through operationalization described in section 5.2.1. Importantly, the intricacies of GLMs, more specifically, the probabilistic nature and inherent uncertainty of GLM outputs, which may manifest in unreliable or hallucinated content, and which our participants seemed to be generally aware of (Memmert, Soroko, & Bittner, 2025), may reinforce this shift, as it makes human curation essential (Memmert, Soroko, & Bittner, 2025). This split

of effort between creation and curation also relates to the calibration of reliance on AI, which is discussed in the literature on AI-assisted decision making (Bućinca et al., 2021; Eckhardt et al., 2026), and which will be an important research area for working with generative AI (Eckhardt et al., 2026). Given that humans integrate GLMs into their everyday work, constantly using them for new tasks, and given that these GLMs evolve, such a calibration should not be considered a one-time act, but an "iterative adaptation" (Richter & Schwabe, 2025, p. 3), in which the human continuously adjusts their effort based on the circumstances (e.g., task and context constraints, comparative ability of themselves and the GLM).

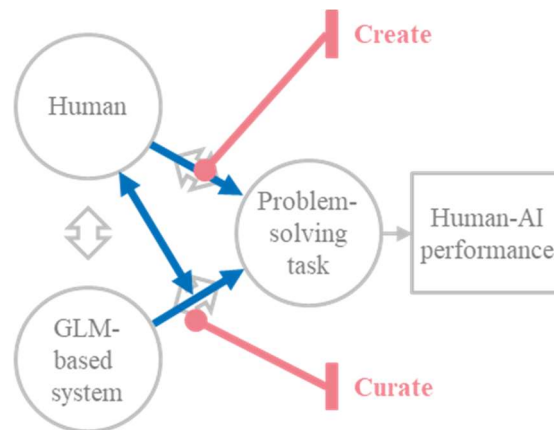


Figure 8. Human creation and curation when working with GLMs

Across studies, neither quantitative measures nor subjective reports indicate a reduction in cognitive load when working with GLMs (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025), contrary to expectations under the smart loafing construct (Stieglitz et al., 2022). The intra-task redistribution of effort may explain why cognitive load does not seem to change (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025): while fewer cognitive resources were required for generating ideas (*create*), more were required for reviewing ideas (*curate*). This is consistent with recent findings in the literature showing that reviewing AI output can be cognitively demanding (Tankelevitch et al., 2024), and with professionals' reports that they carefully check GLM outputs because they are held responsible for the final results (Memmert, Soroko, & Bittner, 2025).

Second, at the **task versus job-level**, the findings suggest a compensatory "fill-up" effect (Memmert, Soroko, & Bittner, 2025). While workers may indeed show reduced visible effort for tasks that they believe the GLM is particularly suitable for, qualitative reports suggest that workers use the saved time to take on additional tasks (Figure 9), which they consider to be of higher value, to increase efficiency at a job level. From this perspective, GLMs do not reduce overall effort; instead, they contribute to a reconfiguration of effort across a broader set of work activities.

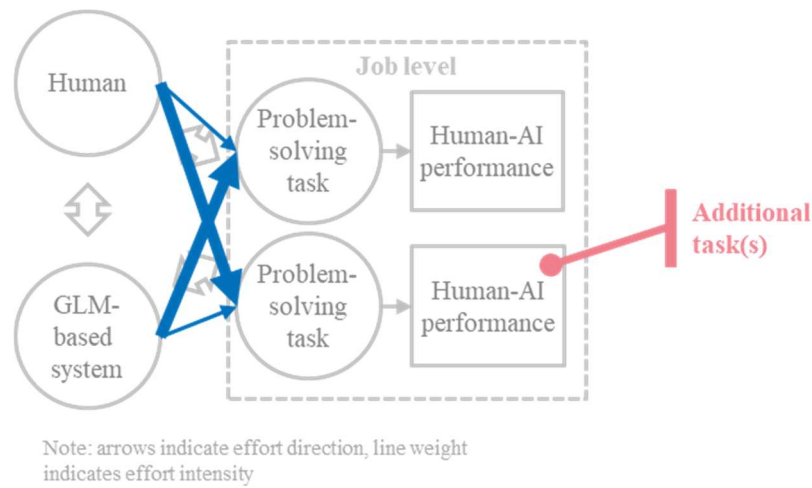


Figure 9. Changed effort intensity and direction across job-level tasks

Third, our findings suggest that **workers initially and ongoingly seek to learn** how they can effectively use GLMs for work, including to understand GLMs' capabilities and limitations, to identify suitable tasks for GLM support, and to integrate GLMs into their workflows (Memmert, Soroko, & Bittner, 2025; Retkowsky et al., 2024). This is additional effort; an effort investment that workers hope will pay off in the future, thus representing a cross-temporal facet of effort management. This initial and continuous learning is required due to the intricacies of GLMs (particular, task agnostic pre-training) and the rapidly evolving model and tool landscape.

In summary, effort management describes how humans deliberately and situationally redistribute effort, more specifically, effort intention (how hard they work on a task) and effort direction (on which task they work) when collaborating with GLMs, due to, inter alia, intricacies of (working with) GLMs, contextual factors like responsibility norms, and job-level efficiency considerations. Effort management builds on social loafing literature (Latané et al., 1979) and the smart loafing concept (Stieglitz et al., 2022), showing that what might be interpreted as reduced (visible) effort (direct human contribution) can mask a reallocation of effort, including learning and experimenting with GLMs, and curating and integrating GLM outputs. This also provides a tentative explanation for the observed stability in cognitive load and aligns with emerging literature on cognitive load when working with GLMs (Tankelevitch et al., 2024). Effort management directly relates to the hybrid intelligence and complementary performance concepts (Dellermann, Ebel, et al., 2019): collective performance improves when the relative strengths of human and AI are combined, for example, when the AI generates and the human evaluates (Memmert, Cvetkovic, et al., 2025), or when the human conceptualizes and the AI operationalizes (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025). Of course, this requires adequate human attention to and engagement with GLM outputs (see Vreede & Memmert, 2025).

The findings have several implications for future research on effort when working with GLMs. First, effort management might be difficult to observe in lab studies with only one

simple task, where humans cannot dynamically reallocate their effort across subtasks or (job) tasks, and the cross-temporal effort investment—such as learning how to effectively use the GLM for this task—is not visible. Secondly, applying existing metrics for effort or social loafing from human groups to human–AI settings can be a good starting point for understanding the effects of GLMs (e.g., visible effort proxied via the number of ideas contributed, cf. Elshan et al., 2025). However, without sufficient contextualization, such measures risk mischaracterizing human–GLM collaboration by overlooking shifts toward less visible but performance-relevant activities, including evaluation and integration (i.e., curation). Moreover, GLMs may reshape the motivational drivers for effort, like expectancy or instrumentality (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025; Vreede & Memmert, 2025). If AI becomes more capable, understanding these shifts and the potential erosion of motivation and effort may become even more important. Lastly, Sarker et al. (2019) call for, in addition to instrumental, considering humanistic goals. In light of our findings of a fill-up effect, GLMs partially taking over work aspects traditionally considered human strengths, and human effort shifting away from creating own toward curating AI contributions, one might ask if we move towards what could be described as a *mechanization of knowledge work and problem-solving*, with uncertain long-term consequences.

5.2.4 Responsibility attribution

Working with others as part of a group can lead to a feeling of diffused responsibility (Pinsonneault et al., 1999). When humans work with a GLM instead of alone, an additional entity—though non-human—enters the picture, forming a human–GLM dyad or group. Humans can socially respond to technical systems (Nass & Moon, 2000) and AI systems may be perceived as teammates (Richter & Schwabe, 2025; Rix, 2022; Seeber et al., 2020). This raises the question of whether humans, when working with AI—similarly to when working with other humans—feel reduced responsibility (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022). For example, Stieglitz et al. (2022) found that individuals attribute some responsibility to AI systems, and Henkenjohann and Trenz (2024) find humans feel less responsible when working with AI. As competence is important for AI to be perceived as a teammate (Rix, 2022), and given that GLMs become increasingly capable, we argue that understanding responsibility diffusion is important because it is *inter alia* a known antecedent of changes in effort and social loafing (Karau & Williams, 1993; Latané et al., 1979).

Through our studies, we investigated responsibility both quantitatively and qualitatively. In our experimental brainstorming studies, after participants had brainstormed jointly with the GLM, we asked them to quantify their feeling of perceived responsibility (Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023). With this approach, we followed prior research on responsibility attribution when working with AI (Stieglitz et al., 2022). We contrasted the felt responsibility with directly observable contributions made by those humans as part of the brainstorming study (Memmert & Tavanapour, 2023). Moreover, we

asked participants to justify their reasoning as to why they feel a certain degree of responsibility, both in our experimental brainstorming studies (Memmert & Tavanapour, 2023), but also the participants in our study who use GLMs for their professional work (Memmert, Soroko, & Bittner, 2025).

We find that across studies (Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023), *most of the participants did not report feeling fully responsible for the results*, despite being the only human working on the task. Instead, most of the participants attributed some of the responsibility to the GLM, which seems to confirm earlier findings by (Stieglitz et al., 2022), who also found such responsibility attribution in human–AI settings.

To better understand responsibility attribution, for each participant, we mapped perceived responsibility against the share of ideas they contributed in their respective brainstorming session to the joint human-GLM pool of ideas (Memmert & Tavanapour, 2023). Felt responsibility and share of contribution do not match in most cases, and we find a tendency to report higher felt responsibility compared to the share of ideas contributed. In a follow-up analysis, we find a strong correlation between the participants' contribution share and the participants' felt responsibility (Memmert, Cvetkovic, et al., 2025). Hence, the extent to which a human contributes directly (here, the number of ideas contributed in the brainstorming session) and the degree to which they feel responsible seem related. However, there appear to be other factors involved, as the correlation only accounts for approximately 50% of the variance.

When asking participants about their reasoning for the reported responsibility, we found distinct justifications (Memmert & Tavanapour, 2023). Unsurprisingly, some participants based their argument on the *degree of their direct contribution to the work results*. For example, some cited the percentage of ideas authored by them compared to the percentage authored by the GLM, reflecting and confirming our quantitative analysis described above (Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023).

Besides mentioning their own or the GLM's direct contribution, a second line of reasoning pertained to the participants' role in producing the results, particularly in requesting, reviewing, selecting, and adjusting ideas (Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023)—in this dissertation, referred to as *curation*. This may partially explain why participants reported a higher share of responsibility compared to the share of their direct contribution. Another potential explanation might be that participants perceived themselves as having contributed more ideas, essentially feeling as if they had generated ideas that were, in fact, contributed by the GLM. This would mirror behavior documented in prior literature, that, in certain group constellations, humans may overattribute ideas to themselves (P. B. Paulus et al., 1993; Stroebe et al., 1992).

Interestingly, when we asked participants who use GLMs for their professional work (Memmert, Soroko, & Bittner, 2025), *participants reported feeling fully responsible for the*

results, a stark contrast to our findings from the experimental brainstorming studies discussed above, and what was suggested in the prior literature (Stieglitz et al., 2022). We surfaced multiple justifications, confirming and extending prior findings. First, some participants did not feel they had to credit the GLM at all, given that they perceived it as merely a tool. Second, in the brainstorming studies, participants had one task for which they could use a GLM by clicking a button, and the GLM would provide them with ideas. In contrast, in the professional use study, participants had to determine when to use a GLM, which part of the task to use it for, formulate prompts, integrate their own contributions with those of the GLM, and then share the work products with their colleagues. Hence, the process administration or curation role seemed to be even more important (also see Retkowsky et al., 2024). Third, participants described that others would hold them responsible or accountable for the results, irrespective of how they were produced, that is, whether they had used a GLM. And, fourth, some reported personal ethical and professional standards they wanted to maintain. In summary, in professional, organizational contexts, responsibility was firmly anchored to the human, due to GLM usage behavior, formal accountability norms, and own standards (Memmert, Soroko, & Bittner, 2025).

Thus, while seemingly confirming earlier findings in literature that human–AI collaboration can lead to diffused responsibility or partial responsibility attribution to the AI (cf. Henkenjohann & Trenz, 2024; Stieglitz et al., 2022), our results show that such attribution may be contingent upon the direct contribution of the human, the way they interact with the GLM and the degree to which they orchestrate the process, and to circumstantial factors like the organizational context and relationships with others (Memmert, Soroko, & Bittner, 2025). Hence, responsibility perceptions may be contingent on the situated task (i.e., moderated by the task and context) and not a stable attribute of human–AI groups.

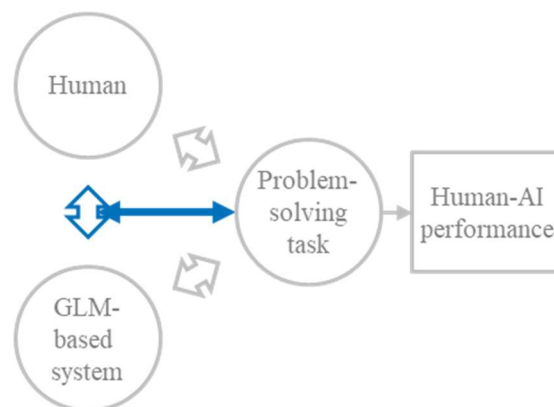


Figure 10. Responsibility attribution between the human and the GLM when working on a situated problem-solving task

Our findings have some implications for studying responsibility attribution and, in the context of GLMs becoming more capable and agentic. Investigating responsibility in experimental settings with participants not feeling the need to comply to professional standards when using GLMs, and without interaction with others (e.g., having to share the results

with a colleague) may only offer limited insight into responsibility (re-)formation when working with GLMs, because the loose or exploratory environment may encourage shared attribution, whereas professional contexts reinforce human accountability (Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023).

Our findings tie into the AI as a tool versus a teammate discussion. First, some participants did not attribute any responsibility to AI, because AI was seen as a tool, whereas others seemed to at least implicitly perceive it as another entity. Future research should investigate whether this perception influences responsibility attribution. Second, participants, besides their direct contribution, refer to the curation or orchestration they perform when integrating GLMs into professional work. The systems in our studies, while able to contribute relevant content, acted at the request of the human, and many professionals stressed the need for manual contribution of their own content and curation of GLM content, resulting in feeling fully responsible. Thus, at least for now, organizational norms and the way GLMs are integrated into professional workflows seem to anchor responsibility with the human (Memmert, Cvetkovic, et al., 2025). However, if systems become more capable or autonomous, requiring less human intervention (e.g., curation or orchestration), perceived responsibility may shift, and humans may attribute more responsibility to the GLM and feel less responsible themselves, which may have negative consequences (Henkenjohann & Trenz, 2024), given that feeling of responsible is an antecedent of reduced effort or social loafing (Pinsonneault et al., 1999). Likewise, given that participants reported that the relationship with the recipients of their work is important when responsibility is concerned (Memmert, Soroko, & Bittner, 2025), and, at least for now, using GLMs for core aspects of work may be perceived negatively by coworkers (Niederhoffer et al., 2025), understanding whether the broad adoption and increasing familiarity with GLMs lead to a gradual shift in the perception of content jointly produced with GLMs, and if this also shifts the perceived responsibility.

5.3 Influencing factors (RQ 3)

Besides understanding the mechanisms that shape the outcomes, it is important to understand the circumstances that may influence these mechanisms (Whetten, 1989). Below, the key elements of the sociotechnical system are discussed, namely, the individual differences of the human, context- and task-related aspects, as well as the intricacies of the technical system. The dissertation specifically focused on problem-solving, which is why, although different tasks were investigated as part of this study—brainstorming, DSR concept development, coding, and professional knowledge work—a task-specific differentiation was out of scope.

5.3.1 Individual: Task-related ability

Prior research called for understanding how characteristics of individuals—such as differences in ability—influence who benefits from technology use. This question is important

in research on ICSS (Holzner et al., 2025; K. Wang & Nickerson, 2017), human–AI collaboration more broadly, and was also discussed in the context of (creatively) using GLMs (Benbya et al., 2024; Dell’Acqua et al., 2023).

In this dissertation, individual ability is considered with respect to the task. For brainstorming tasks, for example, creative ability and topic knowledge are particularly relevant (Memmert, Cvetkovic, & Bittner, 2024), and for conducting DSR, familiarity with the domain and the DSR paradigm is important (Memmert, Tavanapour, & Bittner, 2023). Accordingly, references to 'novices' and 'experts' throughout this dissertation should be understood as relative to the task at hand rather than as general labels. Therefore, the different studies employed different metrics for assessing ability, which reflect the diversity of investigated tasks, but which are not further distinguished here. To that end,

- we explored experts' expectations on the usefulness of GLMs for novices (Memmert, Tavanapour, & Bittner, 2023);
- analyzed the difference in ratings of novices and experts for GLM-produced contents (Memmert, Cvetkovic, & Bittner, 2023);
- analyzed qualitative accounts of participants as to the interaction with and usage of GLM-produced contents (Memmert & Tavanapour, 2023);
- performed an experiment to analyze how ability affects usage behavior and results (Memmert, Mies, & Bittner, 2024).

We deliberately decided to investigate self-reported ability because one's own beliefs can shape behavior and are therefore theoretically relevant (Bandura, 1997). However, we also quantitatively measured the ability to address potential inaccurate self-assessments.

Expected benefits and subjective evaluation. When asking DSR experts about what they thought, who benefits most from GLM-based tool support, they anticipated novices to benefit most, but also stated a potential risk of overreliance (Memmert, Tavanapour, & Bittner, 2023). Somewhat contrary, when we had GLMs generate content for said tool, there was a significant positive correlation between the helpfulness rating and expertise: the higher the expertise, the more helpful the GLM suggestions were rated (Memmert, Cvetkovic, & Bittner, 2023), so expectations about who benefits may differ from personal evaluation.

GLM usage behavior. When considering reliance on GLMs, the participants who feel less capable seem to rely more on the GLM. For example, in our brainstorming studies, participants reported not being knowledgeable about the brainstorming topic or not being creative, and hence made requests for GLM suggestions (Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023). In professional work settings, participants reported using GLMs particularly for tasks that are not at their core expertise (Memmert, Soroko, & Bittner, 2025). This may suggest that perceived ability shapes willingness to use GLM support, which is consistent with the literature (M. Dowsett et al., 2024). When using measured ability (via creativity test), as opposed to self-reported ability, the same pattern occurs.

We find that low-ability individuals request and adopt a higher share of GLM ideas compared to high-ability individuals (Memmert, Mies, & Bittner, 2024), which suggests that actual and perceived ability align in their behavioral effects.

Benefitting from GLMs. Regarding who benefits from GLM usage, low-ability individuals show the strongest performance improvements (Memmert, Mies, & Bittner, 2024), consistent with prior work (e.g., Dell’Acqua et al., 2023; Holzner et al., 2025). There are multiple potential explanations. One possibility is that making more use of GLM suggestions contributes to these gains—our study shows that low-ability individuals request more and adopt a higher share of GLM ideas. However, our data does not allow causal attribution. Another explanation is the larger comparative performance gap: prior work shows that highly creative humans can outperform GLMs in idea generation (e.g., Haase & Hanel, 2023; Holzner et al., 2025; Koivisto & Grassini, 2023), so even with GLM usage, high-ability individuals may have fewer opportunities to improve.

Unintended side effects. As discussed before, GLM use can negatively impact human performance (see section 5.1). In our brainstorming studies, human fluency decreased across ability groups (Memmert, Cvetkovic, et al., 2025; Memmert, Mies, & Bittner, 2024), and also specifically for high-ability participants (Memmert, Mies, & Bittner, 2024). This is particularly relevant given that high-ability individuals tend to contribute the most original human ideas (Haase & Hanel, 2023; Holzner et al., 2025). Moreover, consistent with experts’ concerns about novice overreliance (Memmert, Tavanapour, & Bittner, 2023), we observed indications of such overreliance when novices conducted a coding project (Memmert, Borchers, et al., 2025).

In summary, expectations—experts believe novices benefit most—align with actual behavior, as novices rely more on GLM support (also see Mark Dowsett, 2025). However, expectations and perceptions can diverge: despite anticipating greater benefits for novices, experts personally evaluate GLM content as more helpful (Memmert, Cvetkovic, & Bittner, 2023). Hence, across studies, individual differences play a critical role, shaping GLM usage behavior, performance benefit potential (Dell’Acqua et al., 2023), and side effects such as reduced own performance, positioning individual differences as an important influential factor for human–AI work.

5.3.2 Situated task: Contextual norms and task structure

In the previous sections, effort management (section 5.2.3) and responsibility attribution (section 5.2.4) were discussed as mechanisms. The following section focuses on how and why these mechanisms may manifest differently. While no single study explicitly compared situated tasks, multiple studies have investigated the mechanisms in different situated tasks, allowing for a cautious and exploratory analysis.

Contextual norms and responsibility. The studies we conducted surfaced conflicting reports on responsibility attribution when working with GLMs, as described in detail in section 5.2.4. In our controlled experimental studies on idea generation (Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023), we find that participants do not claim full responsibility for the results but attribute a share of responsibility to the GLM. A follow-up analysis further showed that participants' reported felt responsibility correlated with the proportion of ideas they contributed. However, they tended to report higher responsibility than their actual contribution share would imply. Qualitatively, they explain their reasoning with the share of their own direct contribution, or with the indirect contribution curating the GLM output. Our findings seemingly confirm earlier findings of responsibility attribution to AI systems (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022).

However, this phenomenon of attributing responsibility to the GLM does not seem to be present in other contexts, for example, when humans use GLMs in professional contexts (Memmert, Soroko, & Bittner, 2025). When asked about their perceived responsibility, participants reported feeling fully responsible for the results jointly developed with the GLM—a stark contrast to the experimental findings and to prior literature (see, e.g., Henkenjohann & Trenz, 2024; Stieglitz et al., 2022). Participants mentioned organizational accountability norms (e.g., employees are held accountable for their work), stakeholder relationships (i.e., social relationships with clients and colleagues who have expectations for the work results), and reputational considerations (e.g., using GLMs for core parts of their own work may lead to being perceived as less competent; also cf. Niederhoffer et al., 2025), which collectively anchor responsibility to the human worker.

This suggests that responsibility diffusion is not inherent to human–GLM work, but rather a context-dependent phenomenon. Responsibility tends to diffuse in experimental tasks lacking real-world consequences or social expectations, but remains with the human when organizational norms and interpersonal relationships are present. This finding indicates a boundary condition for responsibility attribution to AI systems.

Task-structure and effort shifts. Echoing the above, our earlier studies on brainstorming with a GLM showed that humans working with a GLM produced fewer ideas compared to solitary brainstormers (Memmert, Cvetkovic, et al., 2025), resembling—on a surface level—a reduction in effort, i.e., social loafing-like behavior, seemingly confirming potential negative side-effects of AI use discussed in the literature (cf. M. Liu et al., 2023; Stieglitz et al., 2022). This impression was further strengthened by qualitative accounts of participants reporting having become "lazy" (Memmert, Cvetkovic, et al., 2025). In contrast, when examining professionals' work practices (Memmert, Soroko, & Bittner, 2025), we uncovered something different: workers *redistributed* effort across their jobs' tasks, taking on higher-level tasks, additional tasks, or investing time in learning to better steer the GLM. Thus, what may appear as effort reduction in an experimental single-task setting may manifest as effort reallocation in a multi-task organizational context. These findings suggest

that the visibility and interpretation of effort mechanisms may depend on contextual factors, particularly whether the context allows for effort to be shifted between tasks. Laboratory studies using a single task, by design, cannot capture such redistribution. Thus, the apparent effort reduction observed in such experiments may partly be a consequence of task structure rather than an inherent response to GLM collaboration.

In summary, contextual factors—particularly organizational norms and social relationships—play a key role in shaping how responsibility and effort mechanisms manifest in human–GLM collaboration. These findings carry important implications for studying human–AI collaboration. First, they emphasize the importance of adopting a sociotechnical perspective: responsibility and effort are not inherent properties of humans or GLMs alone, but rather emerge from their interaction within specific situational tasks (e.g., organizational contexts). Second, the results indicate that directly transferring constructs and measurement approaches from human group research (e.g., social loafing) to human–AI settings (Elshan et al., 2025) may be misleading. Hence, future research should reflect GLM usage behavior in practice—for example, by using multi-task setups and incorporating accountability norms—when investigating how responsibility is attributed, and effort is (re-)distributed.

5.3.3 Technology: Intricacies of GLMs

Besides the individual and the situated task, the technical system is a key element when conceptualizing human–GLM problem-solving as a sociotechnical system. In our studies, we examined GLMs because they possess certain characteristics that distinguish them from prior systems. These characteristics, (1) generativity and adaptivity, and (2) opacity and uncertainty, give rise to and shape the mechanisms described earlier.

Generativity and adaptability. GLMs can *generate*—not just retrieve or classify—content, allowing them to support open-ended problem-solving tasks, such as creative idea generation or concept development, thereby going beyond AI supporting routine decision-making tasks. GLM’s output can introduce new content, enabling the exploration of new directions or perspectives (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert & Tavanapour, 2023). This *content-level* contribution capability also leads to a change in user behavior: besides *creators*, users also act as *curators*, reallocating effort toward checking, selecting, and integrating GLM-generated content (Memmert, Cvetkovic, et al., 2025). Moreover, GLMs can hallucinate or misrepresent facts (Memmert, Borchers, et al., 2025; Memmert, Soroko, & Bittner, 2025), which our participants seemed to be generally aware of (Memmert, Soroko, & Bittner, 2025). However, this also means that humans must invest effort in evaluating and verifying content, such as checking correctness (Memmert, Borchers, et al., 2025), cross-referencing sources (Memmert, Soroko, & Bittner, 2025), or validating whether ideas align with the task context (Memmert & Tavanapour, 2023). Contributing content is a key part of problem-solving, leading some in-

dividuals to attribute partial responsibility to the GLM—at least in some settings (Memmert, Cvetkovic, et al., 2025). However, having to fact-check and integrate GLM contents leads some humans to attribute more or full responsibility to themselves (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025).

GLMs are also *adaptive*, as their outputs can be shaped through prompting and iterative refinement, i.e., GLMs dynamically adapt to user inputs at inference time without requiring retraining. In our brainstorming studies, ideas already mentioned in the same session were provided to the (embedded) GLM as additional input by dynamically including them in the prompt (Memmert, Cvetkovic, et al., 2025; Memmert, Mies, & Bittner, 2024; Memmert & Tavanapour, 2023). Our study on prompt engineering showed the difference prompt design can make for idea creativity (Memmert, Cvetkovic, & Bittner, 2024). Moreover, professionals in our study were able to condition the GLMs' outputs through prompting strategies, refining tone, structure, or target audience (Memmert, Soroko, & Bittner, 2025).

Combining generativity and adaptability, GLMs can provide dynamic, situation-aware support (i.e., a dynamic scaffolding mechanism). However, they also require the user to learn when and how to use the system, and shift effort partially from creation to curation (i.e., effort management; also see Retkowsky et al., 2024). This need for targeted prompting and orchestration is also sometimes cited by users as a rationale for claiming responsibility beyond direct content contribution (i.e., responsibility attribution). Hence, the GLM's capabilities of generativity and adaptivity fundamentally change human–AI work, enabling a form of collaboration that has implications for outcomes and human–AI dynamics.

Novel, evolving, and probabilistic. Besides these key enabling characteristics, our studies have shown that several properties of GLMs, as well as the dynamic nature of how the GLM tool landscape evolves, create challenges for users.

GLMs can be considered a new class of systems that expands which tasks can be supported with AI, but whose task-relevant affordances are not necessarily understood by all users (Benbya et al., 2024). Moreover, using GLMs productively is contingent upon being able to elicit the right content from the GLM. Therefore, users need to invest substantial effort, for example, in learning prompting strategies or discovering suitable use cases for employing GLM (Memmert, Soroko, & Bittner, 2025; Memmert, Tavanapour, & Bittner, 2023). Moreover, the rapidly evolving GLM tool landscape (new versions, new tools) further contributes to this challenge: professionals described the need for constantly catching up (Memmert, Soroko, & Bittner, 2025). Additionally, GLM output can—depending on the GLM's configuration—be probabilistic (i.e., different results for the same user input), which can make it more difficult to form an understanding of the GLM's capabilities and its usefulness for a task, requiring repeated experimentation (Memmert, Soroko, & Bittner, 2025). This creates ongoing, not one-off, learning costs. This is why we frame user behavior as *effort management*, not effort reduction (cf. M. Liu et al., 2023): GLM use incurs effort beyond the task at hand. For researchers, it is important to be aware of this effort,

which cannot be captured by single-task experiments. Managers of workers who use GLMs may also incorporate this when considering (and judging) workers' effort; accounting for effort investment into skill building and workflow adaptation seems important (Memmert, Soroko, & Bittner, 2025; Retkowsky et al., 2024).

In summary, GLMs can be considered a new class of information systems, possessing unique capabilities that enable their use for problem-solving (i.e., dynamic scaffolding and problem- and solution-space exploration), but also exhibit characteristics that alter effort patterns and responsibility attribution.

5.4 Implications for human–GLM design (RQ 4)

The focus of this dissertation lies on understanding the phenomenon and underlying mechanisms of human–GLM problem-solving, which can serve as a foundation for deriving design knowledge (Gregor, 2006). Yet, within the DSR paradigm, providing design knowledge typically requires thorough evaluation (Venable et al., 2016). Accordingly, this section distinguishes between evaluated design contributions and tentative implications derived from explanatory findings. Only two studies (Memmert, Cvetkovic, & Bittner, 2024; Memmert, Tavanapour, & Bittner, 2023) evaluated design knowledge—regarding GLM-based scaffolding and embedded prompt design—and only one developed design principles. Insights from the remaining studies can inform, but not yet prescribe, system design and should therefore be treated as tentative.

5.4.1 GLM-based adaptive support

Particularly novices can struggle when approaching unfamiliar problem domains, such as when students develop a DSR study concept with limited prior experience and largely work on it by themselves. To explore how GLMs might provide effective support in such contexts, two interlinked studies examined the design (Memmert, Tavanapour, & Bittner, 2023) and feasibility (Memmert, Cvetkovic, & Bittner, 2023) of GLM-based adaptive support, covering different aspects of the sociotechnical phenomenon.

The first study (Memmert, Tavanapour, & Bittner, 2023) developed design principles for a tool that combines hard scaffolding—for example, a structured canvas for articulating *elements of a DSR concept* (e.g., issues, design requirements, design principles, kernel theories; see also Meth et al., 2015)—with dynamic scaffolding generated by a GLM. These principles were instantiated in a conceptual user interface and interaction design and evaluated through interviews with DSR experts. Our second study (Memmert, Cvetkovic, & Bittner, 2023) showed that GLMs can generate elements of a DSR concept that are rated as helpful by scholars.

Together, these studies show GLMs' potential to support novices by combining fixed structures with content-level suggestions, helping them, for example, to overcome the blank-page problem or identify relevant literature. However, experts also highlighted the need to

prevent unintended usage patterns. Our contribution includes both a situated instantiation and evaluated design principles, positioning our studies as early, empirically grounded contributions to a broader effort to design GLM-based support systems for research and learning (e.g., Dwivedi et al., 2023; Si et al., 2025; Susarla et al., 2023).

5.4.2 Embedded prompt design

In some GLM-based applications, users do not directly interact with the model through a conversational interface. Instead, the GLM-interaction is abstracted away and integrated into a graphical user interface. Users may, for example, interact with the model through buttons that trigger idea generation or content suggestions; we hence refer to this as "embedded GLM capabilities" (Memmert, Cvetkovic, & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). In such systems, the formulation of the prompt or prompt template becomes a developer-side design decision made at development time. Prompt templates may then be dynamically populated at runtime based on the task context, for example, with the ideas already generated or elements of a DSR concept (see, e.g., Memmert, Cvetkovic, & Bittner, 2023; Memmert & Tavanapour, 2023). Because GLM outputs are strongly conditioned on the prompt, developing the embedded prompt is a critical design choice.

To examine this design space, our study systematically evaluated how different prompt-engineering techniques—and their combinations—shape the creativity of GLM-generated ideas (Memmert, Cvetkovic, & Bittner, 2024). We systematically generated outputs for all technique combinations and had them assessed for creativity, followed by comparative statistical analyses. The findings provide quantitative insights into prompt construction for creative uses, for example, indicating that including evaluation criteria in the prompt can increase the novelty of GLM-generated ideas (also see section 5.2.2).

While not formalized into design principles, our findings offer early, empirically grounded guidance for designing embedded GLM capabilities in ICSS and informed the development of apps for studies in this dissertation (e.g., Memmert, Cvetkovic, et al., 2025; Memmert, Mies, & Bittner, 2024; Memmert & Tavanapour, 2023). Although models have since evolved and the impact of prompt design is model-dependent, subsequent research continues to show substantial effects of prompt design on the creativity of outputs, even with newer GLMs (Zhang et al., 2025). When designing prompts, it is important to consider the dual influence of GLMs, namely, GLMs contributing directly and indirectly (see section 5.1.2). Prior research showed that not all stimuli are equally helpful for performance; if stimuli are too close to the thought process of the human, they may lead to cognitive inertia or fixation (Kohn & Smith, 2011; Lamm & Trommsdorff, 1973), and consequently the human may not explore the full problem- or solution-space. If stimuli are too far away, they can increase cognitive load, which has a negative effect on creative performance (Redifer et al., 2021). Hence, finding a design that manages the distance or relatedness of stimuli or ideas might be a fruitful avenue for future research.

5.4.3 Tentative implications for design

Beyond the evaluated design contributions, the mechanisms identified across studies also suggest design implications. While offering potential guidance, these are not evaluated and should therefore be treated cautiously.

A first implication concerns designing for active human contribution. Several studies suggest that GLMs can induce passivity or overreliance (e.g., Memmert, Borchers, et al., 2025; Memmert & Tavanapour, 2023; Memmert, Tavanapour, & Bittner, 2023), which can be a problem when users have unique task information that the model cannot access (see also Fügener et al., 2021). Systems may therefore incorporate features that prompt users to provide information on the situated task (e.g., goals, constraints, or domain knowledge) before generating suggestions, or detect and signal the lack of critical information. Interaction designs can further discourage one-click acceptance (Buçinca et al., 2021), for example, by deliberately providing only partial information (Memmert, 2024) or by reviewing user-generated content (see Memmert, Borchers, et al., 2025). Relatedly, usage behavior often differs by skill level (Mark Dowsett, 2025; Memmert, Cvetkovic, et al., 2025): novices may be more open to relying on GLMs and may benefit from stronger scaffolding (Memmert, Tavanapour, & Bittner, 2023). Systems could respond to such differences by adapting the degree of scaffolding based on interaction patterns, such as the amount and quality of user-generated input or the extent to which GLM suggestions are accepted without modification (also see Eckhardt et al., 2026).

A second implication pertains to designing for an appropriate sense of responsibility. While findings from our study indicate no responsibility diffusion (Memmert, Soroko, & Bittner, 2025), related work suggests that AI involvement may diffuse responsibility, particularly in contexts with weaker accountability norms (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022). In human groups, transparency about who contributed what can counteract the diffusion of responsibility. Designers could explore features that make human and GLM contributions transparent, for example, through provenance indicators showing which parts of an output originate from the GLM. Such features may be increasingly important as GLM-based tools become more autonomous, though it could be challenging to implement due to the increasingly intertwined nature of human and GLM contributions (Retkowsky et al., 2024).

While these implications are tentative, they illustrate how the identified mechanisms and influencing factors may inform considerations for system design (see Gregor, 2006), providing an initial foundation for future design science research.

6 Theoretical contribution

6.1 Contributions to the conversation in (IS) literature

Research in IS follows a cumulative tradition, in which contributions should be made to active discourses or conversations in the literature (S. Brown, 2024; Gregory & Henfridsson, 2021; Larsen et al., 2019). Through the publications in this dissertation, we contribute to several discourses and calls for research in the literature, underscoring the relevance of our work.

6.1.1 Human–AI collaboration for problem-solving

Two central calls in the literature motivate this dissertation: (1) a focus on human–AI collaboration rather than automation, and (2) a focus on problem-solving rather than decision-making. First, scholars call for investigating AI that collaborates with humans, complementing or augmenting rather than replacing them through automation (Akata et al., 2020; Berente et al., 2021; Krogh, 2018), positioning AI as machine teammates in joint human–AI work (Rix & Hess, 2022; Seeber et al., 2020). The idea is that humans and AI can work together, complementing each other, achieving what was framed as hybrid intelligence (Akata et al., 2020; Dellermann, Ebel, et al., 2019). Second, scholars call for investigating how AI can help to solve *problems* (Berente et al., 2021; Dellermann, Ebel, et al., 2019; Krogh, 2018; Seeber et al., 2020). Instead of repeatable or routine tasks like classification or prediction problems in AI(-assisted) decision-making, this calls for exploring human–AI collaborative problem-solving for open-ended tasks that are not necessarily codifiable and for which there is no (obvious) ground truth (Berente et al., 2021; Dellermann, Ebel, et al., 2019). Scholars suggest that AI could and should participate in multiple steps of problem-solving, including generating, evaluating, and refining ideas or solutions (Krogh, 2018; Seeber et al., 2020). Combined, these research streams articulate a shared call: to understand how humans and AI can *jointly* solve *problems*, rather than AI automating or assisting in (closed-ended) decision-making tasks.

To answer these calls, this dissertation conceptualizes human–AI problem-solving as a sociotechnical system following prior research (Rix & Hess, 2022; Seeber et al., 2020). It explains core concepts (Memmert & Bittner, 2022), for example, human–AI collaboration as reciprocal interaction between a human and an AI system for joint task work, and problem-solving as working on dynamic, open-ended, uncertain tasks that have no ground truth (cf. Dellermann, Ebel, et al., 2019; Quesada et al., 2005). Providing conceptual clarity and construct definition is a critical foundation for theorizing (Fisher et al., 2021; Gregor, 2006; Whetten, 1989). Moreover, the literature review offered an overview of the early research landscape (Memmert & Bittner, 2022).

Within this sociotechnical systems perspective, generative AI—specifically GLMs—are identified as a central enabler of human–AI problem-solving. We deliberately use the term GLM (instead of LLM) across studies to emphasize that their *generative capacity* sets them apart from prior AI systems. Across the dissertation’s studies, we discuss how GLMs combine generativity and adaptivity to dynamically make *content-level contributions*. This marks a departure from earlier AI systems, which may support routine decision-making or provide meta-level guidance rather than generating substantive task content (Memmert, Cvetkovic, et al., 2025; Memmert, Tavanapour, & Bittner, 2023). This distinction is crucial: it specifies *what is new* about GLMs, fulfilling a key step in phenomenon-centered theorizing (Fisher et al., 2021), and helps sharpen the core constructs of the phenomenon (Whetten, 1989). Because GLMs can generate, elaborate, reframe, and evaluate ideas and solution components, they can participate across multiple steps of problem-solving—aligning with Seeber et al.’s (2020) argument that effective machine teammates must be able to participate in multiple stages of the problem-solving process. Moreover, GLMs are probabilistic, which influences human behavior in response. Highlighting these characteristics sets the stage for the subsequent sections, which theorize the mechanisms, dynamics, and performance implications that emerge when humans collaborate with GLMs to solve problems.

The studies specifically investigate problem-solving tasks (cf. Memmert & Bittner, 2022), with characteristics such as open-endedness, lack of ground truth, and a need for exploration and refinement, making them well-suited research contexts. Our studies provide an empirical, multi-method examination of human–GLM problem-solving across these contexts. In “contribut[ing] to the conversation in IS literature” (S. Brown, 2024, p. iv), they extend the empirical foundation of the discourse on how humans and AI systems collaborate beyond decision-support tasks and concretize how using GLMs shapes work—thereby addressing the calls for research identified earlier. Across papers, we surface idiosyncrasies of and recurring patterns in human–AI work. For example, though human–GLM dyads often outperform humans alone, individual human performance does not necessarily increase (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). Similarly, we find that humans tend to reallocate effort rather than simply reduce effort (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025). Additionally, we show that responsibility attribution to AI may be situationally dependent (Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023). This underscores the need for theorizing this new form of human–AI work, and motivates the subsequent contributions on performance in human–AI collaboration (section 6.1.2), mechanisms and group dynamics (section 6.1.3), and task-specific implications such as contributions to ICSS literature (section 6.1.4).

6.1.2 Superior performance

Performance is a central concern in the human–AI collaboration discourse, with scholars calling for a better understanding of how collaboration with AI affects task outcomes (Seeber et al., 2020). More specifically, achieving complementary team performance or

superior performance of the human–AI group compared to the individual is positioned as a core ambition of hybrid intelligence (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021; Vaccaro et al., 2024).

A first contribution lies in further clarifying what constitutes performance in human–AI work. As discussed in section 5.1, the dissertation disentangles *individual performance* (output produced by the human) from *collective performance* (output produced by the human–GLM dyad). This definitional clarification aligns with theory-building requirements to specify and define constructs (Fisher et al., 2021; Gregor, 2006; Whetten, 1989), contributes to an ongoing definitional discourse on human–AI performance (see also Hemmer et al., 2025), and lays the foundation for theorizing how GLMs shape human behavior and performance outcomes.

A second contribution is empirical. Across multiple studies, the findings show that human–GLM dyads can outperform solitary humans in open-ended creative tasks without a ground truth, with the GLM contributing (creative) content (e.g., Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). This extends prior empirical work that focused on showing human–AI superiority in AI-assisted decision-making (e.g., Hemmer et al., 2025). At the same time, the studies reveal performance-related tensions, for example, that (at least directly observable) individual human performance does not necessarily increase when working with a GLM, but may remain stable or even decline; and that subjective perceptions of performance do not necessarily align with quantitatively measured outcomes. Following the ideas that higher-level phenomena can be explained through lower-level phenomena, we thereby contribute towards uncovering new (sub-)phenomena that warrant future research. In that sense, across studies, *statements of relationship*—for example, that working with a GLM tends to enhance collective performance but does not necessarily enhance individual, observable performance—were developed, derived, or tested (see, e.g., Memmert, Soroko, & Bittner, 2025), which is central to theorizing in the sense of specifying relationships among constructs (Gregor, 2006).

Enhancing our understanding of performance addresses core concerns in the literature and lays the groundwork for understanding why these effects occur, thereby motivating a subsequent discussion of mechanisms and dynamics in human–AI work.

6.1.3 Performance mechanisms and human behavior

Understanding how humans and AI systems work together has been highlighted as an important research need (Hemmer et al., 2025; Makarius et al., 2020). Beyond performance outcomes, scholars emphasize the need to uncover the underlying mechanisms and dynamics that shape human–AI collaboration. A related discourse concerns the transferability of human group work knowledge—for instance, whether theories such as those on social loafing or responsibility diffusion meaningfully apply to settings in which one collaborator is an AI system rather than another human (M. Liu et al., 2023; Siemon, 2023; Siemon et al., 2015; Stieglitz et al., 2022).

Moreover, several concerns have been raised regarding unintended or negative side effects of AI use. This includes negative behavioral effects and dynamics such as increased cognitive load, or, related to the prior point, potential effort reduction or responsibility diffusion; but also negative outcome-related effects, such as reduced diversity or homogenization of results (Dell'Acqua et al., 2023; Holzner et al., 2025; Rix & Hess, 2022; Seeber et al., 2020; Stieglitz et al., 2022; Tankelevitch et al., 2024). Together, these discourses highlight the need to explain the inner workings of human–AI problem-solving and to determine whether and how established mechanisms from human-group research can be applied to this new sociotechnical constellation.

As discussed in section 5.2, the studies uncover four mechanisms: dynamic scaffolding, problem- and solution-space exploration, effort management, and responsibility attribution. These mechanisms explain how GLMs shape performance and ways of working on the problem, as well as human–AI dynamics, thereby offering explanatory depth for the performance patterns discussed in section 6.1.2. Identifying such mechanisms fulfills a central requirement of explanation-and-prediction-oriented theorizing (Type IV, Gregor, 2006), namely specifying how constructs relate, essentially contributing towards uncovering the “cogs and wheels” of the phenomenon (Hedström & Ylikoski, 2010, p. 50).

As elaborated in section 5.3, the dissertation further uncovers what can be referred to as boundary conditions shaping the manifestation of these mechanisms. These include individual characteristics such as (creative) ability, system characteristics enabling this new way of working, such as generativity, and task and contextual factors, such as the presence or absence of certain organizational norms or the nature of the task. These boundary conditions specify when and under which circumstances the mechanisms occur (Busse et al., 2017; Gregor, 2006; Whetten, 1989). Together, the mechanisms and boundary conditions offer an empirically supported theoretical account of how GLMs shape both performance and interaction dynamics in human–AI problem-solving.

A key question in the literature is whether mechanisms that occur and are well-documented in human-group settings also occur in human–AI groups (e.g., M. Liu et al., 2023; Siemon, 2023; Siemon et al., 2015). Our studies used theories from human group work—such as on social loafing and cognitive stimulation—as complementary lenses (Fisher et al., 2021) and examined their applicability in human–GLM work (e.g., Memmert, 2024; Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023). The findings show that social loafing does not manifest simply as a reduction in effort, as the classic understanding would suggest (Karau & Williams, 1993; Latané et al., 1979), but rather as a more nuanced shift in how humans allocate effort. This motivates the effort management mechanism developed in this dissertation (Memmert, Soroko, & Bittner, 2025) and refines existing theorizing by showing that effort is often reallocated toward curation or invested into continuous learning rather than reduced. Similarly, like in human groups, responsibility seems to be diffused when working with AI (Henkenjohann & Trenz, 2024; Memmert & Tavanapour, 2023; Stieglitz et al.,

2022), but—as the dissertation shows—such attribution diminishes in organizational or professional contexts (cf. Memmert, Soroko, & Bittner, 2025). Thereby, the dissertation contributes theoretically by explicating boundary conditions and offering extensions to theory (see Busse et al., 2017, and "theory-as-a-discourse view", Larsen et al., 2019, p. 891), for example, on how responsibility attribution unfolds when one collaborator is an AI (also see Stieglitz et al., 2022).

In summary, the findings illustrate that some human-group mechanisms are a good entry point for analyzing human–AI collaboration, but they require adjustment because GLMs are neither humans nor static tools, a point that is conceptually analyzed in Memmert, Cvetkovic, et al. (2025), further clarifying relevant constructs in human–AI work and refining our theoretical understanding of how group work mechanisms apply—and where they need adaptation—to explain human–AI work.

6.1.4 Contribution to individual creativity support systems

Individual creativity support was deliberately chosen as a research context because creative ideation is central to solving (open-ended) problems. Early work in this dissertation (Memmert & Bittner, 2022) already indicated that generative AI could support idea generation, which is contrary to the common framing of creativity as a core *human* comparative strength (Dellermann, Ebel, et al., 2019; Lichtenthaler, 2018). This tension makes the phenomenon theoretically interesting. Our research builds on two streams: first, the longstanding literature on ICSS, which has surfaced different ways of supporting individual creativity (e.g., Althuizen & Reichel, 2016) and established approaches for evaluating tool effectiveness (K. Wang & Nickerson, 2017); and second, the rapidly evolving work on GLMs and creativity (e.g., Haase & Hanel, 2023; Runco, 2023), which explores whether GLMs can be considered creative and how they influence creative processes. Beyond the broad question of whether GLMs can support creativity, this dissertation contributes to specific discourses and calls for research on the implications of GLM use for creative outcomes, the mechanisms through which GLMs shape relevant processes, potential cognitive or behavioral side-effects such as fixation or reduced effort, and the role of individual differences such as creative ability in determining who benefits from GLM-based creativity support.

Conceptual clarification. Again, a first contribution lies in conceptual clarification. The dissertation delineates how GLM-based ICSS differ from prior ICSS approaches (Memmert, Cvetkovic, et al., 2025). Traditional ICSS typically provide meta-level or process-level support—such as guiding individuals through structured ideation methods—whereas GLM-based ICSS provide what we refer to as *content-level* support by generating ready-to-use, task-specific ideas, dynamically taking into account the human's ideas. The dissertation also clarifies how brainstorming with a GLM differs from brainstorming individually

or with another human, thereby defining the phenomenon and sharpening the relevant constructs, in line with theorizing requirements to specify *what a phenomenon is and what it is not* (Gregor, 2006; Whetten, 1989).

Difference in performance logic. Moreover, the dissertation clarifies performance implications. As discussed in section 5.1, disentangling individual from collective performance is essential, particularly for understanding creativity support. Classical stimuli-based ICSS aim to enhance individual performance (K. Wang & Nickerson, 2017), whereas GLM-based ICSS in this dissertation often leave observed individual performance unchanged or reduced, while the human–GLM dyad achieves higher collective performance (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025). This indicates that traditional ICSS and GLM-based ICSS differ in their performance logic.

The dissertation contributes theoretical explanations by identifying mechanisms through which GLMs shape creative work. Problem- and solution-space exploration (sections 5.2.2) explains how GLMs introduce new directions and categories, expanding the search space beyond what individuals would typically explore. This aligns with recent empirical evidence (Specker et al., 2025) and with arguments that GLMs can broaden innovation search (Bouschery et al., 2023). Effort management (sections 5.2.3) explains how individuals re-allocate effort from generating ideas to curating GLM output, shedding light on whether humans exert less effort when using AI in creative work, a concern raised by Krogh (2018). Together, these mechanisms contribute to an ongoing IS discourse, extending ICSS research by showing that GLM-based support not only stimulates the individual (subjectively) but may change where creative work occurs within the human–AI dyad.

Additionally, the dissertation contributes to discussions on performance implications by engaging with concerns that GLMs may reinforce existing views, narrow exploration, or homogenize creative outputs (Dell’Acqua et al., 2023; Holzner et al., 2025; Seeber et al., 2020). Our studies show that GLMs can broaden individuals’ perspectives—in terms of flexibility—while recent literature warns that, at larger scales, GLM-supported outputs may become more homogeneous. This distinction nuances the debate on idea diversity in GLM-supported creativity, confirming findings that GLMs can make individuals’ ideas more diverse (Holzner et al., 2025).

Role of individual differences. A longstanding question in ICSS research is whether less creative or more creative individuals benefit more from creativity support tools (K. Wang & Nickerson, 2017). Recently, similar questions have been raised regarding GLM usage (e.g., Dell’Acqua, 2023). The dissertation contributes to these discourses by showing that individuals with lower creative ability tend to benefit more from GLM-based support than individuals with higher creative ability (Memmert, Mies, & Bittner, 2024). This finding may hint at an individual-level boundary condition that specifies for whom GLM-based ICSS are likely to be most effective, addressing the *who*-dimension of theoretical explanations (Busse et al., 2017; Whetten, 1989).

The dissertation further offers tentative, empirically grounded explanations for this difference. Less creative individuals tend to request more ideas and adopt GLM suggestions more readily (Memmert, Cvetkovic, et al., 2025). This difference in usage behavior may help explain the difference in benefiting from GLM and point to an avenue for further theorizing. Together, these insights extend the discussion on potential boundary conditions for GLM-based support (also see Dell’Acqua et al., 2023) and complement the mechanism-based explanations presented earlier.

Designing GLM-based ICSS. Embedding GLM capabilities into ICSS requires careful design, as at system development time, when designing the prompt template, one must decide how to frame the task, provide context, and structure the interaction with the user. The dissertation provides initial empirical evidence that prompt formulation shapes the novelty and value of generated ideas (Memmert, Cvetkovic, & Bittner, 2024), highlighting that prompt design is not merely a user skill but a core design choice in building GLM-based ICSS. Although these insights are necessarily tentative given the rapid development of GLMs, they contribute early particular guidance on how designers might configure embedded GLM capabilities in GLM-based ICSS, and they contribute to an ongoing discourse on prompt engineering (e.g., Zhang et al., 2025). Towards theorizing for design and action (Gregor, 2006), the study suggests emerging design knowledge for structuring prompts and contextual information. The different prompt techniques were quantitatively evaluated for their *usefulness* (see Gregor, 2006), yet the design knowledge was not formally codified.

6.2 Summary of theoretical contributions

To summarize, this dissertation offers several complementary theoretical contributions. It sharpens key concepts for theorizing the phenomenon of human–GLM problem-solving (e.g., performance) and explicates *what is new* about this phenomenon—namely, the intricacies of working with GLMs and the implications thereof—thereby contributing through construct specification and differentiation (Fisher et al., 2021; Gregor, 2006; Whetten, 1989). Moreover, it explains changes in performance and behavior by identifying underlying mechanisms (e.g., responsibility attribution) and boundary conditions (e.g., organizational responsibility norms), which are central elements of theory development (Busse et al., 2017; Gregor, 2006; Whetten, 1989). To do so, it tests the applicability or transferability of existing mechanisms and theories from human groups to human GLM dyads (e.g., social loafing), and offers suggestions for extending existing concepts on human–AI work (e.g., smart loafing). Moreover, it derives propositions for future research as key aspects of theorizing (Gregor, 2006). Taken together, these contributions work towards a theory for explaining and predicting on human–GLM problem-solving (Type IV, Gregor, 2006), while offering tentative foundations for future theorizing for design and action (Type V, Gregor, 2006).

6.3 Theoretical discussions on an emerging phenomenon

As pointed out by Monteiro et al. (2022, p. i), "compared to relatively stable phenomena that some fields study, a defining aspect of information systems (IS) phenomena is that they are inherently emerging." This observation also holds true for the phenomenon in this dissertation, particularly after the release of ChatGPT in November 2022, with frequent releases of new model versions and the rapid adoption in practice. At the same time, this phenomenon has been extensively explored in the academic literature. Against this backdrop, the section below reflects on the contributions of this dissertation in the context of this emerging and rapidly evolving phenomenon.

Especially for newly emerging phenomena, understanding the new phenomenon, understanding how it is different from prior phenomena, and working towards clarity on the main concepts is important and essential for theorizing (Fisher et al., 2021). With our work, for example, in untangling performance, we contribute to this search for clarity on a key outcome of human–AI collaboration (Hemmer et al., 2025), which remains important, even as GLMs evolve. Similarly, comparing prior and GLM-based ICSS helps to understand how GLMs support work differently, thereby differentiating this phenomenon from prior systems (see Raisch & Fomina, 2025).

The data collection for multiple of the empirical papers included in this dissertation (e.g., Memmert, Cvetkovic, & Bittner, 2023; Memmert & Tavanapour, 2023; Memmert, Tavanapour, & Bittner, 2023) started before the release of ChatGPT in November 2022. Of course, since then, familiarity with GLMs and knowledge of their capabilities have advanced. Earlier works of this dissertation focused on exploring whether the research challenge to support humans in solving problems (beyond decision-making) with AI (Dellermann, Ebel, et al., 2019; Krogh, 2018; Seeber et al., 2020) could be addressed with GLMs.

This included, for example, exploring prompt engineering techniques to increase the creativity of GLM outputs (Memmert, Cvetkovic, & Bittner, 2024). Since then, many papers have examined the creativity of GLMs (Holzner et al., 2025), and models have evolved significantly. Some studies show that earlier prompt engineering techniques might be less effective in some application areas (G. Wang et al., 2025) or showed the sensitivity of prompt engineering effectiveness to model size (P. Liu et al., 2023), rendering the generalizability of our findings to newer models uncertain. However, newer GLMs do not necessarily generate more creative outputs (Holzner et al., 2025), and even with new models, prompt engineering can significantly enhance the output creativity of GLMs (Meincke et al., 2024; Zhang et al., 2025). Besides their contribution to the literature, these studies, which focused on the technical aspects of the phenomenon, provided an important foundation for subsequent publications. For example, informing the prompt design of the purpose-built brainstorming applications with embedded GLM capability used in multiple studies (Memmert & Bittner, 2024; Memmert, Cvetkovic, et al., 2025; Memmert, Mies, & Bittner, 2024; Memmert & Tavanapour, 2023).

Beyond studies focusing on the technical aspects of the phenomenon, most research has concentrated on the interplay of elements within sociotechnical systems and how GLMs influence behavior and performance, which may be less sensitive to changes in GLMs. For example, this dissertation describes how workers manage effort between subtasks, within tasks of their job, and across time when working with GLMs (see section 5.2.3). Even if GLMs improve, unless tasks are fully automated, workers will still have to continue managing their effort in relation to the GLM and will continue having to explore and learn how to integrate these new GLM capabilities into their jobs. As discussed, if systems become more capable, key motivational drivers could be eroded, or, if systems become more autonomous, responsibility might diffuse, rendering our findings and understanding these dynamics even more relevant.

In summary, human–GLM problem-solving is a quickly evolving phenomenon, and while the more technical aspects of this dissertation may be specifically sensitive to evolving GLMs, understanding how GLMs shape human behavior and performance will likely become more important; and these studies help towards deepening our understanding, uncovering new (sub-)phenomena such as tensions between individual and collective performance as a foundation for future research.

7 Practical implications

7.1 Managers and organizations

Based on our findings, managers and organizations need to understand the performance implications of using GLMs, the intricacies of working with GLMs, and the new dynamics GLMs introduce, to actively support the transition toward this new way of working (see also Retkowsky et al., 2024).

While GLMs can increase creative performance and productivity on the dyadic level (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025), there may be potentially negative performance impacts on an organizational level, specifically in areas requiring creativity (Dell’Acqua et al., 2023; Holzner et al., 2025). Hence, organizations should be aware of and actively manage GLM usage in functions potentially susceptible to such negative performance implications.

Our studies highlighted several intricacies of working with GLMs, setting it apart from prior working with other technologies (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025). This includes workers having to initially and—due to the quickly evolving field with model updates and new GLMs being published—continuously learn how to use and integrate GLM-based systems into their work. Even then, working with the tool itself may not always lead to the desired outcomes. Moreover, the workers shift from being pure creators to being creators and curators (Memmert, Soroko, & Bittner, 2025; Retkowsky et al., 2024). Managers can acknowledge and respond to those intricacies by deliberately encouraging workers to try out using GLMs for their work, which includes providing the relevant infrastructure and relevant support structures, for example, offering help for navigating the quickly evolving GLM-tool landscape and the respective affordances (Memmert, Soroko, & Bittner, 2025).

Lastly, our studies indicate that when there exist strong accountability norms and relationships with the recipients of the work results, workers feel a strong sense of responsibility even when working with GLMs (Memmert, Soroko, & Bittner, 2025). However, as pointed out in the literature (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022), in the absence of such norms or relationships, responsibility can already diffuse today. Several developments may increase the likelihood of responsibility diffusion in the future. When using GLMs for professional work becomes commonplace, norms may shift; we already reported that certain professionals shared draft solutions along with the information that the draft is GLM-generated (Memmert, Soroko, & Bittner, 2025). Such behavior, of course, can create additional work for others in the organization (Niederhoffer et al., 2025; Retkowsky et al., 2024). Moreover, when tools get more autonomous or agentic, humans may feel less of a responsibility for jointly created work results. This suggests that setting and communicating clear responsibility norms is and will become even more important.

7.2 Tool designers

Designers of GLM-based systems need to account for the shift in how humans work with GLMs. Across studies, we observed that humans combine creation and curation—they not only generate content but also evaluate, refine, and integrate GLM-produced content (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025). Consequently, systems should be designed to facilitate this dual role. Therefore, tool designers should develop user interfaces and interaction flows that make it easy to inspect, adapt, and integrate GLM outputs while mitigating (potentially undesired) rapid, one-click GLM generation and copying of content.

Extending the above, our studies showed that, while overall enhancing performance, several potentially undesired behaviors occurred when working with GLMs. First, some participants reported using the GLM because they felt—literally—"lazy" (Memmert, Cvetkovic, et al., 2025), underlining the above suggestion to develop interfaces that encourage engagement. Designers may take inspiration from research on calibrating engagement with and reliance on AI suggestions in decision-making tasks (Buçinca et al., 2021; Eckhardt et al., 2026). However, this may negatively affect willingness to work with such a system, so there is a trade-off to be made (Buçinca et al., 2021). Eckhardt et al. (2026, p. 29) offer potential metrics such as "degree of content modification, or response latency", which could be used to design systems that aid users and reach adequate reliance. Second, for some participants, the GLM seems to create a sense of psychological safety (Memmert & Tavanapour, 2023); they let the GLM provide a direction for working on the task. This can be a problem when, as in our studies, the GLM did not receive more information on the task and the goal than the user did. Tool designers may try to actively detect these situations and actively inform users of this risk. Closely related, third, when users felt less task-related competence, they seemed to rely more on the GLM (Mark Dowsett, 2025; Memmert, Cvetkovic, et al., 2025). This can be a problem if the user has unique knowledge inaccessible to the GLM. Hence, tool designers should have their tool actively elicit relevant information about the situated task, and ask users to add their own—potentially unique—insights before providing suggestions, thereby reducing passivity and encouraging active contribution.

In our studies, we observed differences in usage behavior, with novices making higher use of GLMs (Memmert, Mies, & Bittner, 2024), but also being expected to potentially overrely on the tool (Memmert, Tavanapour, & Bittner, 2023). Hence, tool designers should consider developing responsive interfaces based on user ability or behavior when implementing the above suggestions.

Due to their pre-training, GLMs can generate text across tasks and domains. However, users may find it difficult to understand which of their professional tasks they can use GLMs for (Memmert, Soroko, & Bittner, 2025) or may underutilize GLMs, potentially because they are unaware of what GLMs are capable of, for example, evaluating existing

(instead of merely generating new) code (Memmert, Borchers, et al., 2025). Therefore, tool designers may actively communicate and help users navigate affordances.

Although this dissertation emphasizes a sociotechnical perspective, highlighting the interplay between human, technology, and situated tasks, the technical design remains important. More specifically, the *prompt* is a central means to condition the GLM's output. When *embedding* GLM capabilities into tools and abstracting away the interaction with the GLM through a graphical interface with pre-defined interaction options—in the our studies, for example, the "generate ideas"-button—instead of allowing user to directly interact with the GLM, for example, through a chat interface (e.g., ChatGPT), then the tool designer has to develop a prompt or prompt template. Our study shows that prompt design can have significant implications for output quality, in our case, the creativity of the generated ideas (Memmert, Cvetkovic, & Bittner, 2024). While GLMs evolve, new model versions do not necessarily produce more creative ideas (Holzner et al., 2025), and prompt design seems to remain important to improve the raw output quality of the GLM (Zhang et al., 2025). However, echoing our sociotechnical perspective, this improved raw model output quality does not necessarily improve overall results (Buçinca et al., 2021). Hence, designers should systematically experiment with different prompt designs and monitor both the raw output quality and the joint human–AI results.

7.3 Knowledge workers

Problem-solving is a key part of knowledge work. Our findings suggest that GLMs can positively impact performance in various open-ended tasks, indicating that knowledge workers should experiment with GLMs and consider integrating them into their workflows. However, when doing so, they should be aware of the potential negative side effects of GLM usage.

Beyond well-known issues such as hallucinations, which our participants seemed to be largely aware of (Memmert, Soroko, & Bittner, 2025), GLM usage can lead to passivity, where workers predominantly accept GLM suggestions without contributing their own ideas or knowledge (Memmert & Tavanapour, 2023). While such behavior may be efficient in some cases, knowledge workers should reflect on whether they possess unique information about the situated task and the performance expectation that may be inaccessible to the GLM. If so, they should communicate it to the GLM or adjust the results accordingly. Closely related, users should remain aware that the GLM does not necessarily know more about the task than they do; suggestions by the GLM can potentially create a false sense of safety (Memmert, Cvetkovic, et al., 2025) and should not be mistaken for reliable guidance but should be actively questioned.

In addition to immediate risks, knowledge workers should also consider the longer-term implications of their use of GLMs. Whenever they turn to a GLM, they should reflect on their specific goal at this moment: do they only seek to efficiently complete a task, or do

they also want to develop themselves by learning or improving a new skill (see Memmert, Borchers, et al., 2025). De-skilling has been raised as a significant concern when working with AI (e.g., Rafner et al., 2021), and recent studies have pointed to reduced brain activity when using GLMs (Kosmyna et al., 2025). Hence, depending on their goals, workers should learn to use GLMs in a way that helps them to learn or maintain their skills. For example, in our coding study (Memmert, Borchers, et al., 2025), a more productive approach might involve first attempting to solve the problem independently and then using the GLM to let it *review* code and have it provide structural or specific suggestions, instead of copy-and-pasting the code back and forth between the GLM and the interpreter.

Overall, the results from this dissertation suggest that knowledge workers should use GLMs in a self-reflective manner, leveraging the benefits of these tools while remaining actively engaged and aware of their own goals and unique knowledge. Managers and tool designers can support workers by shaping the environment and tools as described above.

8 Limitations and outlook

Each publication included in this dissertation discusses study-specific limitations. Beyond these, the cumulative dissertation as a whole has several overarching limitations, particularly with respect to (1) its scope, (2) its research design and empirical approach, and (3) the breadth of its contribution, which are outlined below.

8.1 Limitations to scope

There are several limitations regarding the defined scope, including the selection of the constellation, technology, situated tasks, outcome dimension, and the change of the phenomenon itself.

This dissertation focuses on human–GLM dyads. This constellation was deliberately chosen due to its practical relevance (Retkowsky et al., 2024) and its suitability for isolating performance implications and human–GLM dynamics. At the same time, researchers also examine settings in which multiple humans work with one or even multiple AI systems or agents (Schwabe et al., 2025; Specker et al., 2025; Zercher et al., 2023). It remains unclear to what extent and how the mechanisms uncovered here materialize in such constellations, though, for example, effort management and problem- and solution space exploration will continue to be important (see, e.g. Nomura et al., 2024).

As the technical part of the phenomenon, this dissertation examines text-based, generative AI systems (GLMs) that were not fine-tuned. This deliberate choice reflects the importance of reciprocity in collaboration (Bedwell et al., 2012): text can be easily iteratively refined with the GLM at a detailed level in a back-and-forth manner. However, the findings may not fully generalize to other modalities or models that do not offer this flexible, fine-grained interaction (e.g., when an entire image needs to be regenerated), or vastly less or more capable models. The dissertation discusses the potential implications of increasingly autonomous and capable tools, such as those for responsibility diffusion or effort management; however, empirical validation remains future work once these systems are developed. Moreover, using non-fine-tuned GLMs may systematically underestimate the performance potential of GLMs.

As for the situated task, this study investigates facets of problem-solving. Problem-solving is vast and multifaceted, and the publications cover only a subset of problem aspects (creativity-related, DSR research, software development, and knowledge worker tasks), and predominantly short-term problem-solving interactions (with only one longitudinal study). While selecting multiple tasks enhances external validity, other tasks, especially high-stakes or more long-term problem-solving tasks, may produce different behavioral patterns. Although some studies have covered professional settings or longer-term problem-solving,

future research should investigate whether the uncovered mechanisms hold, for example, in high-stakes environments (also see section 5.3 for influencing factors).

In terms of outcome, this dissertation places a strong emphasis on task performance. This is consistent with calls in the literature identifying performance as a central outcome of human–AI collaboration (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021), and aligns with the instrumental goals often emphasized in sociotechnical IS research (Sarker et al., 2019). Nonetheless, other outcome dimensions, including "humanistic outcomes", exist and are important (Sarker et al., 2019, p. 696); we discuss, for example, a "fill-up effect" (Memmert, Soroko, & Bittner, 2025), or potential negative implications for personal development and learning (Memmert, Borchers, et al., 2025). Given the emerging literature on potential negative short- or long-term consequences of GLM use, balancing instrumental and humanistic goals seems especially important (also see Sarker et al., 2019).

Lastly, the phenomenon of human–GLM problem solving is not stable: models evolve, adoption rises, user familiarity increases, and work practices change. While this novelty in capability made and makes it a theoretically and practically interesting problem, this dynamism limits the temporal generalizability of some findings. I reflect on the implications in different parts of the dissertation (also see section 6.3), and specifically call for future research to investigate how mechanisms such as effort management or responsibility attribution evolve as GLMs become more common in work structures.

8.2 Limitations in research design and empirical approach

The chosen research design and empirical approach, particularly the use of a closed-source, commercial product, primarily lab studies with student samples, and self-reported data, result in several limitations.

Many of the studies relied on closed-source, commercial GLMs, primarily OpenAI's ChatGPT and the models GPT-3, GPT-3.5, and GPT-4 accessed via API (Application Programming Interface) and interfaced by our purpose-built apps. Prior to the release of ChatGPT, I self-hosted open-source models; however, these did not offer sufficient or comparable performance (compared to GPT-3) for running our empirical studies; therefore, commercial models were used, with careful documentation of used model versions, inference parameters, and prompts (see, e.g., Memmert, Cvetkovic, et al., 2025). For our feasibility-oriented study (Memmert, Cvetkovic, & Bittner, 2023), we selected an example such that training data leakage was unlikely, i.e., we chose a paper as a foundation that was published after the training cutoff for the GLM (live online search was not implemented at that time). Despite careful documentation and reporting of relevant inference parameters, using closed-source systems limits reproducibility, as models change over time, are not transparently documented, or are potentially no longer accessible.

Several studies have used experiments with student samples. While experimental settings increase control, and while these studies were complemented by interviews with researchers and professionals (Memmert, Soroko, & Bittner, 2025; Memmert, Tavanapour, & Bittner, 2023), participants across studies were likely more open to and experienced with using technology and GLMs specifically, which may limit the external validity and generalizability with respect to the broader population.

Several studies use self-reported data from participants. While these are informative—because beliefs shape behavior—they remain vulnerable to biases. Across studies, we employed different ways of triangulation, using behavioral data and usage logs, or asking interviewees for in-depth qualitative explanations. Moreover, collecting subjective reports allowed us to uncover tensions between subjective perceptions and measured performance, offering promising avenues for future research. For example, future work could further investigate effort management using psychometric or behavioral measures to differentiate between active curation and passive acceptance of GLM outputs.

8.3 Limitations of theorizing

Aligned with phenomenon-centered research (Fisher et al., 2021; Gregory & Henfridsson, 2021) and working toward addressing the lack of explanatory knowledge for this newly emerging digital technology, the dissertation focuses on developing an in-depth understanding of the human–GLM problem-solving phenomenon. A clear understanding and explanatory knowledge in the form of constructs, mechanisms, and boundary conditions are essential foundations for developing theory on design (Type V, Gregor, 2006). However, beyond one study that follows the DSR paradigm and develops and evaluates design knowledge in the form of design principles (Memmert, Tavanapour, & Bittner, 2023), the dissertation offers only design implications that were not formally evaluated and are, therefore, framed as tentative. Future work should develop, instantiate, and evaluate specific design interventions, for example, supporting users in effort management, preventing responsibility diffusion, or aiding problem- and solution-space exploration.

8.4 Directions for future research

This dissertation develops explanatory knowledge for human–GLM dyads. Future research could examine how mechanisms identified here—such as effort management and responsibility attribution—generalize to other human–AI collaboration constellations (for an overview, see Zercher et al., 2023), or other technical systems (e.g., more capable or more agentic AI systems).

8.4.1 Human–AI constellations beyond dyads

Given that problem-solving in organizations typically involves human teams, it is essential to understand how these mechanisms play out when a GLM is added to a group rather than used by a single individual (e.g., Specker et al., 2025).

This dissertation suggests that responsibility attribution may be moderated by accountability norms and relationships with the recipients of the work: humans tend to feel responsible when their name is associated with the work result or when they have a direct relationship with the recipients of the result (see 5.2.4). Extending this logic to multi-human constellations raises competing expectations. On the one hand, when individuals work alone with a GLM (our studies), contributions remain largely opaque, because only the human knows how much was contributed by them and how much by the GLM; in group settings, however, individual effort may be visible to other human group members, potentially reducing responsibility attribution to the GLM. On the other hand, increasing group size is known to be related to responsibility diffusion and increasing the likelihood of effort reduction or social loafing (Pinsonneault et al., 1999).

8.4.2 AI autonomy and capability

Moreover, future research may explore whether and how more capable or autonomous systems alter these mechanisms. Studies show that GLMs can already produce more creative ideas than many humans (Haase & Hanel, 2023). In line with that, our studies (Memmert, Cvetkovic, et al., 2025; Memmert, Soroko, & Bittner, 2025; Memmert & Tavanapour, 2023) showed that GLMs may (start to) erode classical drivers for motivation, and, thereby, effort, like the feeling of being able to contribute meaningfully to the results (expectancy), or that the own contribution is necessary (instrumentality). While in our studies, this seemed to have led to a redistribution, rather than a reduction in effort, a more capable (or competent, Rix & Hess, 2022) system may induce such negative effects. Likewise, this study focused on GLMs, which, while generating content that was dynamically conditioned on the human input, did not act proactively and performed (large parts of the) work mostly independently of the human; hence, in professional settings, we did not observe responsibility diffusion (Memmert, Soroko, & Bittner, 2025). However, our results showed that felt responsibility and direct contribution are correlated (Memmert, Cvetkovic, et al., 2025). Hence, if systems become more capable and autonomous, performing larger parts of the task by themselves, independently of humans, responsibility may shift.

8.4.3 Formation of effort management

This dissertation discusses effort management as essential for human–GLM problem solving. In principle, for any given sociotechnical system—human, problem-solving task, and GLM—there should exist an ideal allocation of effort by the human that would maximize performance (or other relevant outcomes). In practice, however, humans may not (always) match this ideal; in our brainstorming study, for example, some participants had only the

GLM generate ideas, others almost exclusively relied on themselves, and others were somewhere in between (Memmert & Tavanapour, 2023). Whether this was an ideal effort distribution for each participant is unknown, and qualitative responses offered only initial potential insights into the reasoning for the usage behavior. In a different study, some professionals reported that working with the GLM sometimes feels like a time sink (Memmert, Soroko, & Bittner, 2025). This raises questions about how effort management (strategies) are formed, how far they deviate from optimal effort allocation, and how effort management changes with experience. Future research could therefore examine how humans learn when and for which tasks to use GLMs, as well as how long and to what extent to engage with them. This could shed light on whether effort management strategies exist, if they converge toward more effective patterns over time, or if and how negative patterns form (see M. Liu et al., 2023). Moreover, researchers could explore how systems should be designed to support users in developing more appropriate effort management strategies.

9 Conclusion

GLMs are being rapidly adopted in industry, and while they can substantially improve performance, they may also introduce risks. This dissertation positions GLMs as enablers, with their *generative* capacity allowing them to support on a *content-level*, dynamically generating ideas, text, or code when working with a human. This makes human–GLM problem-solving a timely and theoretically interesting sociotechnical phenomenon, as it fundamentally changes for which kinds of tasks, and how, AI can support humans.

This dissertation investigates how working with GLMs shapes both performance outcomes and human behavior. It shows that mechanisms known from human group work can serve as a useful starting point for analyzing human–GLM work, but that they require careful contextualization to avoid misleading conclusions. In particular, GLMs can alter how effort is allocated, how responsibility is perceived, or how humans approach problem-solving tasks. By developing explanatory knowledge around these mechanisms, this dissertation contributes to the ongoing discourse on human–AI collaboration performance (Bansal et al., 2021; Dellermann, Ebel, et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021) and human–AI dynamics (Makarius et al., 2020), and offers specific insights into the use of GLMs for individual creative support (K. Wang & Nickerson, 2017).

This dissertation represents an early step in understanding a rapidly evolving phenomenon, with new models and model versions being released and humans becoming more familiar with integrating them into their work practices. As models become more capable or more autonomous, questions around human effort (allocation) or responsibility attribution will likely become even more relevant for productive human–AI problem-solving.

References

- Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8), 18–28.
- Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of Management Information Systems*, 33(1), 11–44. <https://doi.org/10.1080/07421222.2016.1172439>
- Alvesson, M., & Sandberg, J. (2011). Generating Research Questions Through Problem-ization. *Academy of Management Review*, 36(2), 247–271. <https://doi.org/10.5465/amr.2009.0188>
- Amabile, T. M. (1988). A model of creativity and innovation in organizations. *Research in Organizational Behavior*, 10(1), 123–167.
- Avital, M., & Te'eni, D. (2009). From generative fit to generative capacity: exploring an emerging dimension of information systems design and task performance. *Information Systems Journal*, 19(4), 345–367. <https://doi.org/10.1111/j.1365-2575.2007.00291.x>
- Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, 45(1), 315–341. <https://doi.org/10.25300/MISQ/2021/15882>
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. W H Freeman.
- Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1). <https://doi.org/10.1007/s12525-023-00680-1>
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *CHI '21, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445717>
- Baruah, J., & Paulus, P. B. (2016). The Role of Time and Category Relatedness in Electronic Brainstorming. *Small Group Research*, 47(3), 333–342. <https://doi.org/10.1177/1046496416642296>
- Bedwell, W. L., Wildman, J. L., DiazGranados, D., Salazar, M., Kramer, W. S., & Salas, E. (2012). Collaboration at work: An integrative multilevel conceptualization. *Human Resource Management Review*, 22(2), 128–145. <https://doi.org/10.1016/j.hrmr.2011.11.007>

- Benbya, H., Strich, F., & Tamm, T. (2024). Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work. *Journal of the Association for Information Systems*, 25(1), 23–36. <https://doi.org/10.17705/1jais.00861>
- Benz, C., Riefle, L., & Satzger, G. (2024). User Engagement and Beyond: A Conceptual Framework for Engagement in Information Systems Research. *Communications of the Association for Information Systems*, 54, 331–359. <https://doi.org/10.17705/1CAIS.05412>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing Artificial Intelligence. *MIS Quarterly*, 45(3), 1433–1450. <https://doi.org/10.25300/MISQ/2021/16274>
- Bouschery, S. G., Blazevic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2), 139–153. <https://doi.org/10.1111/jpim.12656>
- Brown, S. (2024). Editor's Comments: Maximizing the Chances of Securing a Revision by Focusing on the Essentials. *MIS Quarterly*, 48(3), iii–viii. <https://doi.org/10.25300/MISQ/2024/483E1>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Busse, C., Kach, A. P., & Wagner, S. M. (2017). Boundary Conditions: What They Are, How to Explore Them, Why We Need Them, and When to Consider Them. *Organizational Research Methods*, 20(4), 574–609. <https://doi.org/10.1177/1094428116641191>
- Chakraborty, A., Tayob, Y., & Rao, B. (2024). *Reinventing Enterprise Operations*. Accenture. <https://www.accenture.com/content/dam/accenture/final/accenture-com/document-3/Accenture-Reinventing-Enterprise-Operations-FA-9-25-24.pdf>
- Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025). *The GenAI Divide: State of AI in Business 2025*. MIT NANDA.
- Coskun, H., Paulus, P. B., Brown, V., & Sherwood, J. J. (2000). Cognitive stimulation and problem presentation in idea-generating groups. *Group Dynamics: Theory, Research, and Practice*, 4(4), 307–329. <https://doi.org/10.1037/1089-2699.4.4.307>
- Dean, D., Hender, J., Rodgers, T., & Santanen, E. (2006). Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, 7(10), 646–699. <https://doi.org/10.17705/1jais.00106>

- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.
- Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In T. Bui (Chair), *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-41089-5>
- Dörner, D., & Funke, J. (2017). Complex Problem Solving: What It Is and What It Is Not. *Frontiers in Psychology*, 8, 1153. <https://doi.org/10.3389/fpsyg.2017.01153>
- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. <https://doi.org/10.1126/sciadv.adn5290>
- Dowsett, M [M.], Maiden, N., & van Werven, R. (2024). *Epistemic Flux: Creative Professionals, Identity and the Equivocality of GenAI*. <https://openaccess.city.ac.uk/id/eprint/33503/>
- Dowsett, M [Mark] (2025). The Proximity of Expertise: Designers, Everyday Work and the Affordances of Generative AI. In *Academy of Management Proceedings*. Symposium conducted at the meeting of Academy of Management.
- Dugosh, K. L., Paulus, P. B [P. B.], Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *J. Pers. Soc. Psychol.*, 79(5), 722–735. <https://doi.org/10.1037/0022-3514.79.5.722>
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). “So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71. <https://doi.org/10.1016/j.ijinfo-mgt.2023.102642>
- Eckhardt, S., Kühl, N., Dolata, M., & Schwabe, G. (2026). A Survey of AI Reliance. *ACM Computing Surveys*, 58(6), 1–37. <https://doi.org/10.1145/3776528>

- Edmonds, L. (2025). 3 years old, 800 million users, 29,000 prompts a second: ChatGPT's meteoric rise, by the numbers. *Business Insider*. <https://www.businessinsider.com/chatgpt-by-the-numbers-2025-11>
- Elshan, E., Vreede, T. de, Vreede, G. de, Ebel, P. A., & Siemon, D. (2025). Idle Minds: The Role of Social Loafing in Human-AI Teams. In *Proceedings of the 58th Hawaii International Conference on System Sciences*.
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Fisher, G., Mayer, K., & Morris, S. (2021). From the Editors—Phenomenon-Based Theorizing. *Academy of Management Review*, 46(4), 631–639. <https://doi.org/10.5465/amr.2021.0320>
- Frensch, P. A., & Funke, J. (1995). *Complex Problem Solving: The European Perspective*. Taylor and Francis.
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, 45(3), 1527–1556. <https://doi.org/10.25300/MISQ/2021/16553>
- Gau, M., Maedche, A., & vom Brocke, J. (2022). DSR Buddy: A Conversational Agent Supporting Design Science Research Activities. *Wirtschaftsinformatik 2022 Proceedings*, 35. https://aisel.aisnet.org/wi2022/prototype_track/prototype_track/35/
- Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3), 611–642. <https://doi.org/10.2307/25148742>
- Gregory, R. W., & Henfridsson, O. (2021). Bridging Art and Science: Phenomenon-Driven Theorizing. *Journal of the Association for Information Systems*, 22(6), 1509–1523. <https://doi.org/10.17705/1jais.00703>
- Guilford, J. P., Merrifield, P. R., Christensen, P. R., & Frick, J. W. (1961). Some New Symbolic Factors of Cognition and Convergent Production. *Educational and Psychological Measurement*, 21(3), 515–541. <https://doi.org/10.1177/001316446102100301>
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Hamza, M., Siemon, D., Akbar, M. A., & Rahman, T. (2024). Human-AI Collaboration in Software Engineering: Lessons Learned from a Hands-On Workshop. In *7th ACM/IEEE International Workshop on Software-intensive Business* (pp. 7–14). ACM. <https://doi.org/10.1145/3643690.3648236>
- Hedström, P., & Ylikoski, P. (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, 36(1), 49–67. <https://doi.org/10.1146/annurev.soc.012809.102632>
- Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2025). Complementarity in human-AI collaboration: concept, sources, and evidence. *European Journal*

- of *Information Systems*, 34(6), 979–1002.
<https://doi.org/10.1080/0960085X.2025.2475962>
- Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In *25th Pacific Asia Conference on Information Systems*, Virtual Event / Dubai, UAE.
<https://aisel.aisnet.org/pacis2021/78>
- Henkenjohann, R., & Trenz, M. (2024). Challenges in collaboration with Generative AI: Interaction patterns, outcome quality and perceived responsibility. In *ECIS 2024 Proceedings*, Paphos, Cyprus.
- Holzner, N., Maier, S., & Feuerriegel, S. (2025). Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis. <https://arxiv.org/abs/2505.17241v1>
- Hüffmeier, J., Dietrich, H., & Hertel, G. (2013). Effort Intentions in Teams: Effects of Task Type and Teammate Performance. *Small Group Research*, 44(1), 62–88.
<https://doi.org/10.1177/1046496412472242>
- Jansson, D. G., & Smith, S. M. (1991). Design Fixation. *Design Studies*, 12(1), 3–11.
[https://doi.org/10.1016/0142-694X\(91\)90003-F](https://doi.org/10.1016/0142-694X(91)90003-F)
- Jiang, Y., & Nakatani, K. (2025). Exploring Implementations of GenAI in Teaching IS Subjects and Student Perceptions. *Journal of Information Systems Education*, 36(2), 180–194. <https://doi.org/10.62273/WFHO1011>
- Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706.
<https://doi.org/10.1037/0022-3514.65.4.681>
- Kim, M. C., & Hannafin, M. J. (2011). Scaffolding Problem Solving in Technology-Enhanced Learning Environments (TELEs): Bridging Research and Theory with Practice. *Computers & Education*, 56(2), 403–417.
<https://doi.org/10.1016/j.compedu.2010.08.024>
- Kohn, N. W., & Smith, S. M. (2011). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 25(3), 359–371.
<https://doi.org/10.1002/acp.1699>
- Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1), 13601.
<https://doi.org/10.1038/s41598-023-40858-3>
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task.
<https://arxiv.org/abs/2506.08872v1>
- Krogh, G. von (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404–409.
<https://doi.org/10.5465/amd.2018.0084>

- Lamm, H., & Trommsdorff, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. *European Journal of Social Psychology*, 3(4), 361–388. <https://doi.org/10.1002/ejsp.2420030402>
- Larsen, K. R., Hovorka, D., Dennis, A., & West, J. D. (2019). Understanding the Elephant: The Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles. *Journal of the Association for Information Systems*, 20(7), 887–927. <https://doi.org/10.17705/1jais.00556>
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.*, 37(6), 822–832. <https://doi.org/10.1037/0022-3514.37.6.822>
- Lichtenthaler, U. (2018). Substitute or Synthesis: The Interplay between Human and Artificial Intelligence. *Research-Technology Management*, 61(5), 12–14. <https://doi.org/10.1080/08956308.2018.1495962>
- Liu, M., Ke, W., & Xu, D. J. (2023). Will Humans be Free-Riders? The Effects of Expectations for AI on Human-AI Team Performance. *PACIS 2023 Proceedings*.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9), 1–35. <https://doi.org/10.1145/3560815>
- Maaravi, Y., Heller, B., Shoham, Y., Mohar, S., & Deutsch, B. (2021). Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science*, 15(6), 1431–1464. <https://doi.org/10.1007/s11846-020-00400-5>
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). Ai-Based Digital Assistants. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>
- Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. <https://doi.org/10.1016/j.jbusres.2020.07.045>
- Meincke, L., Mollick, E. R., & Terwiesch, C. (2024). *Prompting Diverse Ideas: Increasing AI Idea Variance*. <https://doi.org/10.48550/arXiv.2402.01727>
- Memmert, L. (2024). Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects. In *Proceedings of the 32nd European Conference on Information (ECIS)*, Paphos, Cyprus. https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1
- Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>

- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., Borchers, M., Plückhahn, J., & Bittner, E. (2025). Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy. In *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, HI, USA. <https://hdl.handle.net/10125/109714>
- Memmert, L., Cvetkovic, I., & Bittner, E. (2023). Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/405
- Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/107289>
- Memmert, L., Cvetkovic, I., Tavanapour, N., & Bittner, E. (2025). Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load. *Business & Information Systems Engineering*. Advance online publication. <https://doi.org/10.1007/s12599-025-00974-y>
- Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. <https://aisel.aisnet.org/icis2024/aibus/aibus/7/>
- Memmert, L., Soroko, D., & Bittner, E. (2025). From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI. *Business & Information Systems Engineering*(67), 615–635. <https://doi.org/10.1007/s12599-025-00960-4>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
- Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. *Journal of Information Systems Education*, 34(3), 279–292. <https://aisel.aisnet.org/jise/vol34/iss3/3>

- Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the Association for Information Systems*, 16(9), 799–837. <https://doi.org/10.17705/1jais.00408>
- Mingers, J. (2001). Combining IS Research Methods: Towards a Pluralist Methodology. *Information Systems Research*, 12(3), 240–259. <https://doi.org/10.1287/isre.12.3.240.9709>
- Monteiro, E., Constantinides, P., Scott, S., Shaikh, M., & Burton-Jones, A. (2022). Editor's Comments: Qualitative Research Methods in Information Systems: A Call for Phenomenon-Focused Problematization. *MIS Quarterly*, 46(4), i–xviii. <https://doi.org/10.25300/MISQ/2022/464E1>
- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Niederhoffer, K., Kellerman, G. R., Lee, A., Liebscher, A., Rapuano, K., & Hancock, J. T. (2025). AI-Generated “Workslop” Is Destroying Productivity. *Harvard Business Review*.
- Nijstad, B. A., Dreu, C. K. W. de, Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1), 34–77. <https://doi.org/10.1080/10463281003765323>
- Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, 10(3), 186–213. https://doi.org/10.1207/s15327957pspr1003_1
- NISO CRediT Working Group. *ANSI/NISO Z39.104-2022, CRediT, Contributor Roles Taxonomy*. NISO.
- Nomura, M., Ito, T., & Ding, S. (2024). Towards Collaborative Brainstorming among Humans and AI Agents: An Implementation of the IBIS-based Brainstorming Support System with Multiple AI Agents. In *ACM Collective Intelligence Conference Proceedings*. ACM. <https://doi.org/10.1145/3643562.3672609>
- Oliver, L. J. (1982). Helping students overcome writer's block. *Journal of Reading*, 26(2), 162–168. <https://www.jstor.org/stable/40029248>
- Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.
- Paulus, P. (2000). Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied Psychology*, 49(2), 237–262. <https://doi.org/10.1111/1464-0597.00013>
- Paulus, P. B., Dzindolet, M. T., Poletes, G., & Camacho, L. M. (1993). Perception of Performance in Group Brainstorming: The Illusion of Group Productivity. *Personality and Social Psychology Bulletin*, 19(1), 78–89. <https://doi.org/10.1177/0146167293191009>

- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Popovici, M.-D. (2024). ChatGPT in the Classroom. Exploring Its Potential and Limitations in a Functional Programming Course. *International Journal of Human–Computer Interaction*, 40(22), 7743–7754. <https://doi.org/10.1080/10447318.2023.2269006>
- Pyöriä, P. (2005). The concept of knowledge work revisited. *Journal of Knowledge Management*, 9(3), 116–127. <https://doi.org/10.1108/13673270510602818>
- Quesada, J., Kintsch, W., & Gomez, E. (2005). Complex problem-solving: a field in search of a definition? *Theoretical Issues in Ergonomics Science*, 6(1), 5–33. <https://doi.org/10.1080/14639220512331311553>
- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A Scaffolding Design Framework for Software to Support Science Inquiry. *Journal of the Learning Sciences*, 13(3), 337–386. https://doi.org/10.1207/s15327809jls1303_4
- Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., & Sherson, J. (2021). Deskill, Upskill, and Reskill: a Case for Hybrid Intelligence. *Morals & Machines*, 1(2), 24–39. <https://doi.org/10.5771/2747-5174-2021-2-24>
- Rai, A., Constantinides, P., & Sarker, S [Saonee] (2019). Editor’s comments: Next-generation digital platforms: Toward human-AI hybrids. *MIS Quarterly*, 43(1), iii–ix. <https://doi.org/10.25300/MISQ/2019/431E0>
- Raisch, S., & Fomina, K. (2025). Combining Human and Artificial Intelligence: Hybrid Problem-Solving in Organizations. *Academy of Management Review*, 50(2), 441–464. <https://doi.org/10.5465/amr.2021.0421>
- Redifer, J. L., Bae, C. L., & Zhao, Q. (2021). Self-efficacy and performance feedback: Impacts on cognitive load during creative thinking. *Learning and Instruction*, 71. <https://doi.org/10.1016/j.learninstruc.2020.101395>
- Reiser, B. J. (2004). Scaffolding Complex Learning: The Mechanisms of Structuring and Problematising Student Work. *Journal of the Learning Sciences*, 13(3), 273–304. https://doi.org/10.1207/s15327809jls1303_2
- Retkowsky, J., Hafermalz, E., & Huysman, M. (2024). Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. *Business Horizons*, 67(5), 511–523. <https://doi.org/10.1016/j.bushor.2024.04.009>
- Richter, A., & Schwabe, G. (2025). “There is No ‘AI’ in ‘TEAM’! Or is there?” – Towards meaningful human-AI collaboration. *Australasian Journal of Information Systems*, 29. <https://doi.org/10.3127/ajis.v29.5753>
- Rix, J. (2022). From tools to teammates: Conceptualizing humans’ perception of machines as teammates with a systematic literature review. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.

- Rix, J., & Hess, T. (2022). Hello, Mate! Insights from the Field on Leveraging Machine Teammates in Organizations. In *PACIS 2022 Proceedings*. <https://aisel.aisnet.org/pacis2022/95>
- Runco, M. A. (2023). Updating the Standard Definition of Creativity to Account for the Artificial Creativity of AI. *Creativity Research Journal*, 1–5. <https://doi.org/10.1080/10400419.2023.2257977>
- Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, 24(1), 92–96. <https://doi.org/10.1080/10400419.2012.650092>
- Sarkar, A. (2023). Enough With “Human-AI Collaboration”. In A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, & A. Peters (Eds.), *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–8). ACM. <https://doi.org/10.1145/3544549.3582735>
- Sarker, S [Suprateek], Chatterjee, S., Xiao, X., & Elbanna, A. (2019). The Sociotechnical Axis of Cohesion for the is Discipline: Its Historical Legacy and its Continued Relevance1. *MIS Quarterly*, 43(3), 695–719. <https://doi.org/10.25300/MISQ/2019/13747>
- Saye, J. W., & Brush, T. (2002). Scaffolding Critical Reasoning About History and Social Issues in Multimedia-supported Learning Environments. *Educational Technology Research and Development*, 50(3), 77–96. <https://doi.org/10.1007/BF02505026>
- Schaschek, M., Spatscheck, N., & Winkelmann, A. (2024). For Those about to Rely—a Taxonomy of Experimental Studies on AI Reliance. In *19th International Conference on Wirtschaftsinformatik*, Würzburg, Germany.
- Schwabe, G., Katsiuba, D., Specker, R., & Dolata, M. (2025). AI-ThinkLets for Brainstorming. In *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, HI, USA.
- Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2). <https://doi.org/10.1016/j.im.2019.103174>
- Si, C., Hashimoto, T., & Yang, D. (2025). *The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas*. <https://doi.org/10.48550/arXiv.2506.20803>
- Siemon, D. (2023). Let the computer evaluate your idea: evaluation apprehension in human-computer collaboration. *Behaviour & Information Technology*, 42(5), 459–477. <https://doi.org/10.1080/0144929X.2021.2023638>
- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.
- Siemon, D., & Wank, F. (2021). Collaboration With AI-Based Teammates – Evaluation of the Social Loafing Effect. In *PACIS 2021 Proceedings*. <https://aisel.aisnet.org/pacis2021/146/>

- Specker, R. J., Bucher, A., Katsiuba, D., Dolata, M., & Schwabe, G. (2025). An Extra Brain In The Room: Enhancing Large Group Brainstorming With An Autonomous AI-Agent. In *ECIS 2025 Proceedings*. https://aisel.aisnet.org/ecis2025/human_ai/human_ai/15/
- Stieglitz, S., Mirbabaie, M., Möllmann, N. R. J., & Rzycki, J. (2022). Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. *Information Systems Frontiers*, 24(3), 745–770. <https://doi.org/10.1007/s10796-021-10201-0>
- Stroebe, W., Diehl, M., & Abakoumkin, G. (1992). The Illusion of Group Effectivity. *Personality and Social Psychology Bulletin*, 18(5), 643–650. <https://doi.org/10.1177/0146167292185015>
- Sun, R., & Deng, X. (2025). Using Generative AI to Enhance Experiential Learning: An Exploratory Study of ChatGPT Use by University Students. *Journal of Information Systems Education*, 36(1), 53–64. <https://doi.org/10.62273/ZLUM4022>
- Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S [Suprateek] (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, 34(2), 399–408. <https://doi.org/10.1287/isre.2023.ed.v34.n2>
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The Metacognitive Demands and Opportunities of Generative AI. In *CHI '24 Proceedings*.
- Thuan, N. H., & Antunes, P. (2022). Positioning Design Science as an Educational Tool for Innovation and Problem Solving. *Communications of the Association for Information Systems*, 51. <https://aisel.aisnet.org/cais/vol51/iss1/1>
- Treffinger, D., Isaksen, S., & Stead-Dorval, K. B. (2003). Creative Problem Solving (CPS Version 6.1™) A Contemporary Framework for Managing Change.
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, 25(1), 77–89. <https://doi.org/10.1057/ejis.2014.36>
- Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the Qualitative–Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. *MIS Quarterly*, 37(1), 21–54. <https://doi.org/10.25300/MISQ/2013/37.1.02>
- Vreede, T. de, & Memmert, L. (2025). AI-Assisted Task Performance: The Moderating Role of Assistance Modality on Motivational Dynamics. In *Proceedings of the 46th International Conference on Information Systems (ICIS)*, Nashville, Tennessee, US. https://aisel.aisnet.org/icis2025/user_behav/user_behav/22
- Wang, G., Sun, Z., Ye, S., Gong, Z., Chen, Y., Zhao, Y., Liang, Q., & Hao, D. (2025). Do advanced language models eliminate the need for prompt engineering in software

- engineering? *ACM Transactions on Software Engineering and Methodology*, Article 3771933. Advance online publication. <https://doi.org/10.1145/3771933>
- Wang, K., & Nickerson, J. V. (2017). A literature review on individual creativity support systems. *Computers in Human Behavior*, 74, 139–151. <https://doi.org/10.1016/j.chb.2017.04.035>
- Wang, K., & Nickerson, J. V. (2019). A Wikipedia-based Method to Support Creative Idea Generation: The Role of Stimulus Relatedness. *Journal of Management Information Systems*, 36(4), 1284–1312. <https://doi.org/10.1080/07421222.2019.1661095>
- Whetten, D. A. (1989). What Constitutes a Theoretical Contribution? *Academy of Management Review*, 14(4), 490–495. <https://doi.org/10.2307/258554>
- Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- Zercher, D., Jussupow, E., & Heinzl, A. (2023). When AI joins the Team: A Literature Review on Intragroup Processes and their Effect on Team Performance in Team-AI Collaboration. In *ECIS 2023 Proceedings*. https://aisel.aisnet.org/ecis2023_rp/307
- Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., & Shi, W. (2025, October 1). *Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity*. <https://doi.org/10.48550/arXiv.2510.01171>

10 Publication 1: Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts

Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>

Abstract

Solving complex problems has been proclaimed as one major challenge for hybrid teams of humans and artificial intelligence (AI) systems. Human-AI collaboration brings immense opportunities in these complex tasks, in which humans struggle, but full automation is also impossible. Understanding and designing human-AI collaboration for complex problem solving is a wicked and multifaceted research problem itself. We contribute to this emergent field by reviewing to what extent existing research on instantiated human-AI collaboration already addresses this challenge. After clarifying the two key concepts (complex problem solving and human-AI collaboration), we perform a systematic literature review. We extract research contexts and assess them considering different complexity features. We thereby provide an overview of existing and guidance for designing new, suitable research contexts for studying complex problem solving through human-AI collaboration and present an outlook for further work on this research challenge.

10.1 Introduction

Solving complex problems has been proclaimed as one major challenge for hybrid teams of humans and artificial intelligent (AI) systems (Dellermann, Ebel, et al., 2019). The nature of work changes towards higher knowledge intensity and complexity, often exceeding the abilities of individual humans (Jacobs, 2017). In parallel, AI, in particular machine learning, has improved over the last years even outperforming humans in some areas. However, those AI systems are limited to specific, narrowly defined tasks, far from reaching a general level of intelligence (Grace et al., 2018).

Dellermann, Calma, et al. (2019) describe a combination of human and AI – named “hybrid intelligence” by different authors (Akata et al., 2020; Dellermann, Calma, et al., 2019) – as a likely paradigm for the next years. The underlying reasoning is that the strengths and weaknesses of humans and AI systems are complementary, resulting in a mutually beneficial relationship when combined, enabling the team to jointly achieve more than individually possible (Dellermann, Calma, et al., 2019). According to Seeber et al. (2020) for AI-enabled machines to be “effective teammates” they need to “engage in at least some of the steps in a complex problem solving process”. However, how humans and AI systems collaborate to solve problems is still unclear (Benbya et al., 2021).

Though the challenge of complex problem solving (CPS) through human-AI collaboration was posed comparatively recently (Dellermann, Ebel, et al., 2019), there is already an extensive body of knowledge available around the broader topic of human-AI collaboration. However, it is unclear, to what extent this research deals with CPS.

Addressing this open point, we review extant literature on instantiated human-AI collaboration and thereby answer the following research questions (RQ):

RQ 1: In which instantiated research contexts is re-search on human-AI collaboration conducted?

RQ 2: To what extent can these research contexts be considered to deal with complex problem solving?

Seeber et al. (2020) advocate for taking a socio-technical perspective on human-AI collaboration. Figure 1 illustrates a schematic view of socio-technical systems, consisting of *user(s)* interacting with *technology* to work on a *task* within a specific *context*. To what extent a study can be considered to deal with CPS depends on certain characteristics of the *task* that the human-AI team performs and the *context* this task is performed in. Therefore, the combination of *task* and *context* will be the focus of this study. It will be referred to as *research contexts* for the remainder of this paper (some authors use *scenario* as a term, e.g., Frensch & Funke, 1995; Quesada et al., 2005). For the analysis, the research contexts will be extracted from the relevant publications, described briefly (RQ 1) and assessed against literature-based characteristics for CPS (RQ 2).

For researchers wanting to study CPS through human-AI collaboration it is crucial to select an appropriate research context, as the choice of the research context will define the class of the problem the derived insights can be generalized to. This paper provides researchers with an overview of potential research contexts and relevant underlying characteristics.

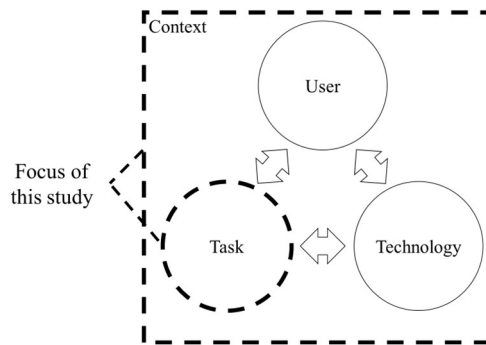


Figure 1. Schematic depiction of a socio-technical system (according to Maedche et al., 2019); focus added

The analysis is limited to studies going beyond only providing theoretical descriptions of potential research contexts, focusing on studies with instantiated research contexts, i.e., actual implementations of research contexts within which user studies or experiments were conducted. This focus ensures that only research contexts are included which have already proven to be suitable for studying CPS through human-AI collaboration. They are therefore thought to be most helpful for scholars as a starting point, both for conducting studies

within these already established research contexts and for inspiring the construction of new research contexts.

Additionally, this focus on instantiations was set in the light of the design science research tradition in IS research (Gregor & Hevner, 2013) and the call for design knowledge (Dellermann, Ebel, et al., 2019; Seeber et al., 2020), with instantiations playing a critical role in theory building (e.g., for developing cause-effect-theories). As Gregor and Hevner (2013) note, “[m]id-range theories or grand design theories” are usually not simply invited, but are built through abstraction from a number of specific instantiations.

The remainder of this paper is structured as follows: As part of the background section, “human-AI collaboration” and its adjacent concepts as well as “CPS” are explained. Within the methodology section the approach for reviewing the literature is described. The overview regarding research contexts and their complexities along with additional explanatory details is provided in the succeeding section. Subsequently, limitations are described and the paper closes with a conclusion and an outlook on relevant future research.

10.2 Background

Both concepts relevant in the context of this paper (human-AI collaboration, CPS) have a long research history outside Information System (IS) research and are described in detail in the following section.

10.2.1 (Human-AI) collaboration

Human-AI collaboration is the idea of humans and AI systems working together. The underlying reasoning is that humans and AI systems are complementary in their strengths and weaknesses (Akata et al., 2020; Dellermann, Ebel, et al., 2019). AI systems can perform clearly defined tasks, process large amounts of data, recognize patterns and make coherent predictions (Dellermann, Ebel, et al., 2019). Already today AI systems can outperform humans who can only process a limited amount of information and are subject to cognitive bias (Akata et al., 2020). Humans, on the other hand, are creative, empathic and have common sense (Dellermann, Ebel, et al., 2019). They comparatively easily navigate dynamic environments, put unexpected events into context and react flexibly to changing circumstances (Akata et al., 2020; Dellermann, Ebel, et al., 2019). They can deal with fragmented information and solve abstract problems (Maedche et al., 2019). Dellermann, Ebel, et al. (2019) provide an in-depth discussion and conceptualization of “hybrid intelligence”.

While there are already many contributions around this topic, there is some uncertainty regarding the terminology. Several papers use the term collaboration without a shared definition (Akata et al., 2020; e.g., Dellermann, Calma, et al., 2019). Additionally, the idea of combining humans and AI systems is researched under a variety of labels, such as human-AI collaboration, human-AI teaming or hybrid intelligence. However, the concepts are not clearly distinguished, which might hinder the progress of research. This terminological

challenge is not only faced by IS research but exists cross-disciplinarily (Bedwell et al., 2012). Bedwell et al. (2012) point out that “without shared understanding of the collaboration construct, it is difficult to theorize and empirically test [...] outcomes associated with collaboration” and therefore state that “it is necessary to reduce construct confusion about collaboration to systematically and collectively advance knowledge”.

Bedwell et al. (2012) provide a comprehensive, multi-discipline literature review identifying the relevant aspects of collaboration, going beyond the idea of *working together* and defining collaboration as an “evolving process whereby two or more social entities actively and reciprocally engage in joint activities aimed at achieving at least one shared goal” (Bedwell et al., 2012). This widely cited review is used as a foundation for the understanding of “collaboration” in this study. The relevant characteristics of the definition are briefly explained according to (Bedwell et al., 2012):

Evolving process. Collaboration is neither structure nor outcome, but a process leading to a (collaborative) outcome. This *process* consists of emergent states (e.g., motivations or values of team members) and collaborative behaviors (e.g., adaptation or sense making) and is influenced by contextual factors (e.g., task or environment). These influences feed back into the collaborative process making it dynamic or *evolving* in nature.

Two or more social entities. Collaboration or “working together” requires at least two entities such as individuals, teams or organizations.

Reciprocal. Collaboration requires mutual engagement of the involved entities, contributing to the collaborative effort, i.e., collaboration is neither one-sided nor simply a delegation of work, but is a “back-and-forth” *reciprocal* process. The requirement for a “back-and-forth” reciprocal process is interpreted in this paper as follows: a simple way of working together with an AI system, such as the AI making a recommendation and a human taking the decision (i.e., decision support system) does not qualify as *collaboration*. The reciprocity characteristic here is in alignment with the understanding of hybrid intelligence where “activities conducted by each agent are conditionally dependent” (Dellermann, Ebel, et al., 2019).

Participation in joint activities. Collaboration focusses on performing joint activities such as solving problems, performing actions, or executing tasks.

Achieving a shared goal. Entities participating in shared work might have a variety of goals, which might be conflicting. However, what sets collaboration apart, is, that all entities have at least one shared or mutually agreed upon goal.

According to the analysis of Bedwell et al. (2012), teaming and collaboration differ to the extent, that collaboration might occur on different levels (e.g., individuals, teams or organizations), whereas teaming usually applies only on team-level (see Figure 2). Given this relationship, both concepts will be considered in this paper as part of the literature search.

Coordination and cooperation lack some of the before mentioned features and are therefore not included (though considered during initial search).

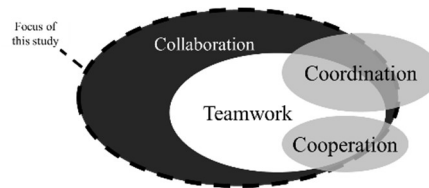


Figure 2. “Shared criterion space among collaboration and related constructs” according to Bedwell et al. (2012); focus added

A widely used means to analyze team effectiveness is the Input-Mediator-Output-Model (I-M-O-Model) (Mathieu et al., 2008). According to Mathieu et al. (2008), inputs can be considered “antecedent factors that enable and constrain members’ interactions”, mediators are processes and emergent states impacting outputs, and outputs are “results and by-products of team activity that are valued by one or more constituencies” (Mathieu et al., 2008). While an in-depth analysis of team effectiveness is beyond the scope of this paper, we will use the model to point out relevant aspects of the studies with regards to CPS through human-AI collaboration.

10.2.2 Complex problem solving

Solving complex problems was proclaimed as the next challenge (Dellermann, Ebel, et al., 2019) for the collaboration of humans and AI systems. Seeber et al. (2020) ask for AI systems in order to be “effective teammates” to “*engage* in at least some of the steps in a complex problem solving *process* [emphasis added]” and Dellermann, Ebel, et al. (2019) outline their understanding of complex problems as being “time variant, dynamic, requir[ing] much domain knowledge and hav[ing] no specific ground truth”. They state that “these highly uncertain contexts require intuitive and analytic abilities, as well as human strengths such as creativity and empathy”.

These descriptions provide a first impression of what could be studied under the label “CPS”. In order to systematically review literature, however, a more detailed understanding of CPS rooted in theory is required.

CPS has a long research history in the psychology literature dating back to the mid-1970s (Dörner & Funke, 2017). It started with researching relatively simple problems in laboratory environments (Frensch & Funke, 1995). However, researchers realized that insights derived from those studies did not generalize well to more complex, real-life problems (Frensch & Funke, 1995). In reaction, European and North American scholars pursued different avenues, with European scholars focusing on increasingly complex, computerized laboratory tasks (so called “microworlds”) whereas Northern American researchers focused on separate, natural knowledge domains (e.g., reading, writing, calculation) (Frensch & Funke, 1995). As a result, key terms such as “complexity” or “CPS” itself are not defined consistently in the field (Frensch & Funke, 1995; Quesada et al., 2005).

Moreover, the transition from *simple* problem solving to *complex* problem solving is not clearly defined either (Dörner & Funke, 2017). Frensch and Funke (1995) state, that these two concepts are qualitatively different. While *simple* problem solving requires to “overcome a single barrier; [*complex* problem solving], in contrast, deals with a large number of barriers that coexist simultaneously. Because there are multiple barriers, a single cognitive or behavioral activity may not be sufficient to reach the goal state”. Problems solvable by *varying one thing at a time* (VOTAT) should not be considered complex problem solving (Dörner & Funke, 2017).

Reflecting on this discourse, Quesada et al. (2005) provide a taxonomy of CPS tasks in order to enable an assessment of “the similarity between tasks and to generalize results from one type of task to another”. The taxonomy contains both “formal” and “psychological” features. Using psychological features for the analysis would require “[postulating] processes and representations [...] in participants facing the task” (Quesada et al., 2005). Given the novelty of the research challenge and the lack of understanding on both processes and representations of joint human-AI CPS in research, only “formal” features are used for the analysis.

The taxonomy consists of three feature-groups (time related, variable related, and system behavior related), each containing three features. Each feature is shortly explained below according to Quesada et al. (2005):

Time related aspects are:

- *Time variance.* While static systems only change when participants intervene, dynamic or time variant systems can change independently of the participant, increasing complexity.
- *Time continuity.* Progression of time might be continuous, e.g., as in most real-life settings or might occur stepwise, e.g., as in games where participants take turns marking discrete “steps”.
- *Degree of time pressure.* Time pressure refers to whether the research contexts require rapid decisions and actions, which increases complexity.

Variable related aspects are:

- *Number and type (discrete/continuous) of variables.* Provides a first indication for the size of the research context or problem, with more variables increasing complexity.
- *Number and pattern of relationships between variables.* Refers to the number of connections between variables. The more interconnected variables are, i.e., the more variables change as a consequence of a manipulation of a variable, the more complex. Certain patterns make the detection of causalities more difficult and therefore increase complexity.

- *Non-linearity*. Refers to the type of connection between variables. As humans do not perform well at estimating non-linear relations, those are considered to increase complexity.

System behavior related aspects are:

- *Opacity*. Refers to whether the complete state of the problem is visible or transparent to the participant. A system containing variables not directly observable (opaque) is considered more complex.
- *Stochasticity*. Refers to the changes within the system resulting from an action. If the same action is taken in the same state always results in the same succeeding state, the system is deterministic. If the resulting state might vary (stochastic), this increases complexity. Some opaque research contexts might appear to be stochastic to participants (because the participants do not see the “full picture”), however, during this literature review the research contexts are assessed with an outside view.
- *Delayed feedback*. Refers to the possibility of relating feedback to actions. If feedback (e.g., successful completion or increased score) cannot be related directly to an action, this increases complexity.

In CPS literature, typically a human is assessed regarding their problem solving competency and strategies. For this paper, we considered a team of at least one human and an AI system. Considering multiple (including non-human) stakeholders makes a clarification regarding *opacity* and *stochasticity* necessary.

Crucial for determining opacity for this study is the combined perception of the participant team, i.e., if the team members can jointly perceive the complete state of the research context, it is not considered to be opaque.

Additionally, stochasticity here is considered a problem characteristic. The fact that one of the team members (e.g., a human that is part of the human-AI team) might act stochastically, does not make the problem or research context stochastic. Simply speaking, a *problem* should not be considered *complex* because a human and AI work on it together but because the problem is inherently complex. Complexity introduced through working with AI systems is out-of-scope for this study as it was discussed already by Yang et al. (2020).

This conceptualization of complexity is actor-independent, which is in alignment with CPS literature: “In the literature on CPS, it is mostly the structure of the external problem representation that is considered complex. So a problem usually is considered being of a certain complexity, even if it might seem less complex to problem solvers with more expertise” (Fischer et al., 2012).

It should be noted, that none of the features are necessary or sufficient to qualify a research context as CPS (Quesada et al., 2005). However, they can be used to compare research

contexts and help to understand, which findings might generalize across which research contexts (Quesada et al., 2005).

10.3 Methodology

In order to ensure transparency and reproducibility, a systematic literature review following vom Brocke et al. (2009) is conducted. The search string used is *"hybrid intelligence" OR ("human" AND ("AI" OR "artificial intelligence") AND ("team*" OR "collaboration" OR "coordination" OR "cooperation"))*. The asterisk works as a wildcard for any number of characters providing flexibility. For databases without wildcard-support the terms *team*, *teams*, and *teaming* were used.

In alignment with Dellermann, Ebel, et al. (2019) *hybrid intelligence* is used to find contributions describing combinations of humans and AI systems. However, as this is a relatively new term for describing human-AI collaboration, contributions containing *human* and *artificial intelligence* (or abbreviated) either in a *collaboration*- or *team*-setting are included additionally. These two concepts (*collaboration*, *team**) are chosen due to their similarity (Bedwell et al., 2012). We presumed that some authors might use the concepts interchangeably and after initial screening we included the terms *coordination* and *cooperation* into our search.

The search string does not contain any reference to CPS as problems or tasks considered to qualify as CPS vary widely (Quesada et al., 2005) and including specific aspects might lead to premature exclusion of relevant search results.

The search string is applied to title, abstracts and keywords. Where possible, the search is restricted to peer-reviewed contributions. Relevant publications for further analysis are identified in two review cycles.

1st review: By examining title, abstract and keywords, the following type of contributions are excluded: “hybrid intelligence” or “AI” have a different meaning than intended in the search string (e.g., “hybrid intelligence” refers to different types of AI algorithms being used together), purely conceptual/ non-instantiated (e.g., theoretical frameworks, use case descriptions), focus on technical aspects of improving machine learning, non-collaborative (in alignment with the understanding outlined earlier), no usage of AI (e.g., suggestion of AI usage only as an outlook), and certain formats (e.g., workshops, talks).

2nd review: By examining the (user) study and research context descriptions within the contributions, only those providing sufficient information to assess the “CPS” characteristics outlined previously are included. This requires a specific description of the problem sought to be solved, not some (more generic) outlook as to which research contexts could potentially benefit from the research. It also requires an instantiation of the research context, e.g., accompanied by a user study in order to maintain the focus on suitable research contexts

and provide a strong foundation for theorizing. Publications with duplicate research contexts (e.g., same context investigated in different publications) and primarily physical tasks (i.e., non-intellectual, Dellermann, Ebel, et al., 2019) were excluded.

The searched domain-relevant databases, the number of results during the initial search, as well as after each of the review cycles are shown in Table 1. We performed a forward and backward search to identify additional relevant articles.

Table 1. Search results per review stage and database, and forward/backward (f/b) search

Database	Initial hits	After 1st review	After 2nd review + f/b
AISel	19	1	-
ACM DL	201	53	16 + 3
ScienceDirect	265	23	2 + 1
Total	485	77	18 + 4

10.4 Results

A total of 22 relevant publications were identified. They were analyzed in order to extract the research contexts used to study human-AI collaboration (RQ 1). Similar research contexts were grouped into research context groups for further analysis. In order to understand to what extent these contexts can be classified as CPS, the CPS taxonomy of Bedwell et al. (2012) was applied (RQ 2).

10.4.1 Research contexts of human-AI collaboration (RQ 1)

Each of the identified publications describes an instantiation of human-AI collaboration in a specific *research context*. Similar research contexts are grouped for reasons of clarity. Table 2 provides an overview regarding the research context groups, the identified examples and the references. We also report exemplary key inputs, mediators and outcomes (IMO) as well as the key findings for the studies, in order to provide readers with inspiration and to facilitate the analysis.

Table 2. Research context groups with examples and exemplary study details

Research context groups/Example(s)	Key Outcome(s)	Key Input(s)/ Mediator(s)	Key Finding(s)
Creative content creation <i>Sketching (Davis et al., 2017; C. Zhang et al., 2021), composing (McCormack et al., 2019; Suh et al., 2021), writing (Calderwood et al., 2020; Yanardag et al., 2021)</i>	<i>PO: Novelty, integrity, interestingness & balance (McCormack et al., 2019; C. Zhang et al., 2021), scariness (Yanardag et al., 2021)</i> <i>PB: Social & collaboration dynamics, flow (Davis et al., 2017; McCormack et al., 2019; Suh et al., 2021)</i> <i>AP: Match writing style, novelty, creativity (Calderwood et al., 2020)</i>	<i>I-TC: AI vs. WoZ (Davis et al., 2017)</i> <i>I-TC: AI vs. no-AI (Suh et al., 2021; C. Zhang et al., 2021)</i> <i>I-TC: AI vs. human-AI vs. human (Yanardag et al., 2021)</i> <i>M-TP: Communication: inner state (McCormack et al., 2019), suggestion style (Calderwood et al., 2020)</i>	<i>Positive effect: Visualization of machine confidence on flow & composition (McCormack et al., 2019)</i> <i>Positive effect: AI on novelty (C. Zhang et al., 2021) and social dynamics (Suh et al., 2021)</i> <i>Positive effect: Hybrid approach on scariness (Yanardag et al., 2021)</i> <i>Writer prefer more fine-grained control over outputs and editability (Calderwood et al., 2020)</i> <i>Inconclusive: Group comparison; focus on framework (Davis et al., 2017)</i>
Turn-based cooperative games <i>Card game (Hanabi), puzzle, word guessing (Ashktorab et al., 2020; Liang et al., 2019; Tabrez et al., 2019)</i>	<i>PO: Score, win/loose, efficiency (Ashktorab et al., 2020; Liang et al., 2019)</i> <i>AP: Helpfulness, intelligence, sociability, humanness, likability, creativity, trustworthiness (Ashktorab et al., 2020; Liang et al., 2019; Tabrez et al., 2019)</i>	<i>I-TC: AI vs. presumed-human (Ashktorab et al., 2020)</i> <i>M-TP: Communication: implicature & explanations (Liang et al., 2019; Tabrez et al., 2019)</i>	<i>Positive effect: Implicature on score and perceived humanness (Liang et al., 2019)</i> <i>Positive effect: Explanation style on helpfulness, trustworthiness and overall experience (Tabrez et al., 2019)</i> <i>Negative attitude towards AI despite equal performance (Ashktorab et al., 2020)</i>
Real-time cooperative games & simulations <i>Dearth, capture the gunner, defend the pass, don't starve together, reconnaissance missions overcooked (Gao et al., 2020; McGee et al., 2011; Merritt & McGee, 2012; Ong et al., 2012; Tulk et al., 2018; Wang et al., 2016)</i>	<i>PO: Score (McGee et al., 2011), efficiency (Gao et al., 2020), mission success (Wang et al., 2016)</i> <i>PB: Protecting or sacrificing teammate (Merritt & McGee, 2012; Ong et al., 2012)</i> <i>AP: Player identity (humanness) (Tulk et al., 2018), perceived helpfulness, efficiency (Gao et al., 2020), trust, understandability (Wang et al., 2016)</i>	<i>I-TC: AI vs. human (McGee et al., 2011; Tulk et al., 2018)</i> <i>I-TC: AI vs. presumed-human (Merritt & McGee, 2012; Ong et al., 2012)</i> <i>I-E: Score visible (Ong et al., 2012)</i> <i>M-TP: Communication: different modalities (McGee et al., 2011)/ explanation types (Gao et al., 2020; Wang et al., 2016)/ skill, reliability (Wang et al., 2016)</i>	<i>Dependability: Visibility of score on treatment of AI teammate (Ong et al., 2012)</i> <i>Humans act differently in presumed social context (Merritt & McGee, 2012)</i> <i>Online players value co-presence (McGee et al., 2011)</i> <i>Positive effect: Staying close to other player on correct identification (AI or human) (Tulk et al., 2018)</i> <i>Positive effect: explanation on non-experts' performance (Gao et al., 2020)/on performance if facilitating decision-making (Wang et al., 2016)</i>

Research context groups/Example(s)	Key Outcome(s)	Key Input(s)/ Mediator(s)	Key Finding(s)
Design <i>Mood board-, bridge-, drone-, accessory-, game level-design (Guzdial et al., 2019; Koch et al., 2020; Lin et al., 2020; Song et al., 2021; G. Zhang et al., 2021)</i>	<i>PO</i> : Strength mass ratio (G. Zhang et al., 2021), quality, novelty (Lin et al., 2020), range, velocity, payload, cost (Song et al., 2021) <i>PB</i> : Design effort, search strategy, mental workload (Song et al., 2021), solution exploration (Koch et al., 2020) <i>AP</i> : Usefulness, inspirational (Koch et al., 2020), fun, frustration, aid, creativity, adaptation (Guzdial et al., 2019), friendliness, engagement (Lin et al., 2020)	<i>I-TC</i> : AI vs. no-AI (Song et al., 2021; G. Zhang et al., 2021) <i>I-E</i> : Task complexity (Song et al., 2021) <i>M-TP</i> : task-related algorithms (Guzdial et al., 2019), embodiment (Lin et al., 2020), agency variants (different features) (Koch et al., 2020)	<i>Positive effect</i> : design on low performing teams (negative on high performing teams) (G. Zhang et al., 2021) <i>Dependability</i> Agency preference of designer and design step; AI algorithm for outcome (Koch et al., 2020) <i>Positive effect</i> : embodiment on satisfaction, friendliness, engagement & helpfulness (Lin et al., 2020) <i>Dependability</i> Algorithm preference on designer (Guzdial et al., 2019) <i>Positive effect</i> : AI on efficiency of solution space exploration (Song et al., 2021) <i>Impact</i> : Task complexity on AI systems effects (Song et al., 2021)
Negotiation <i>Baking items (Geraghty et al., 2020)</i>	<i>PO</i> : Profit <i>PB</i> : Likeability/trust	<i>M-TP</i> : Agents variants – negotiation tactics	<i>Positive effect</i> : Specific negotiation tactics on outcomes and likeability
Investigation <i>Illegal trade (Whitehead et al., 2021)</i>	<i>PO</i> : Efficiency	<i>I-TC</i> : AI vs. no-AI	<i>Presumed positive effect</i> : Hybrid approach on efficiency

PO = performance outcome, PB = performance behavior, AP = agent performance, I-TC = input – team composition, I-E: input – environment, M-TP = mediator – team process; WoZ = Wizard of Oz

The following describes the research context groups along with information regarding the reasoning for using AI in this research context.

Creative content creation. Several studies deal with the creation of creative content (e.g., composing music or drawing a picture). Usually, a human performs an initial creative action (e.g., draws first lines of an idea onto a digital canvas) and can then evoke the AI system to add to the composition. Authors root their argument for choosing AI systems in their research context in literature citing several benefits from previous studies, e.g., C. Zhang et al. (2021) cite the effective provision of inspiration, stimulation of divergent thinking and support for innovation.

Turn-based cooperative games. There is a variety of cooperative games (e.g., card game or puzzles). Usually it is two players taking actions in alternating order. While the puzzles game (Tabrez et al., 2019) allows simultaneous action by both players, it does not necessitate simultaneous action-taking and is thus categorized as turn-based here. The reasoning for choosing AI systems is less related to their abilities within this context but more to the research interest for mediators like social perception (Ashktorab et al., 2020), effective communication (Liang et al., 2019) or explanations (Tabrez et al., 2019).

Real-time cooperative games & simulations. There are several real-time (video) games requiring players to move around and perform actions swiftly. As will be described in the next section, these two types of games (i.e., turn-based, real-time) differ in their complexity and are thus sorted into different research context groups. Similarly to turn-based games, research in the real-time research contexts did primarily focus on collaboration aspects e.g., on how beliefs about teammates manifest in actions (Merritt & McGee, 2012) or behavior (Ong et al., 2012). Due to the similarity, simulations are included in this group.

Design. Five studies are related to design teams supported by AI systems, e.g., focusing on the creation of mood boards (“visual collages composed of images, text, and objects, that express concepts, ideas and emotions”, Koch et al., 2020) or on engineers designing bridges (G. Zhang et al., 2021) or drones (Song et al., 2021). While the humans are in the lead, the AI system provides suggestion for further ideas or improvements. AI systems were chosen to support in those research contexts due to their high task-specific performance expected to be effective when combined with human creativity and agility (G. Zhang et al., 2021) and their ability to support the entire creative process including inspiration and reflection (Koch et al., 2020).

Negotiation. In one study, two sellers (two AI systems) competitively negotiate with a human buyer regarding a transaction. The human buyer and AI system seller might be seen as collaborators with the shared goal of reaching an agreement, making them better off. With each of them aiming to maximize their own utility, they have conflicting goals. This, however, is common within collaborative settings (Bedwell et al., 2012). The research context was chosen due to potential benefits for “retail, e-commerce, legal, business, and industrial sectors” (Geraghty et al., 2020).

Investigation. In one study, an AI system jointly with a team of conservation scientists, law enforcement and criminologists iteratively analyze data to identify instances of illegal trade of plants. AI is used to improve the effectiveness of search and analysis (Whitehead et al., 2021).

In summary, most of the human-AI collaboration research contexts are either real-life creative content creation or design tasks, or well-defined game environments (both turn-based and real-time). For the former, the AI systems generative ability is supposed to provide inspiration to humans supporting the creative process. In many cases, performance with and without AI is investigated (input – team composition). For the latter, the focus is usually not set on game-specific aspects but on researching aspects of collaboration such as effective communication, focusing more on mediating aspects.

Reflecting on the characteristics of collaboration described earlier, an *evolving process* and *reciprocity* are essential. All these research context require a certain degree of sense-making, i.e., trying to understand the situation as well as actions and intentions of the other(s) and adjusting accordingly (e.g., match the intentioned drawing style, Davis et al., 2017). They are also inherently iterative and interdependent, requiring a back-and-forth between

agents. Games and simulations allow researchers more fine-grained control over the environment enabling them to enforce those characteristics, e.g., by creating hard dependencies through providing agents with unique capabilities or restricting access to information, making communication necessary to achieve the joint goal (van den Bosch et al., 2019). This is more difficult for real-world tasks (e.g., writing, drawing, composing).

Measuring team effectiveness is a way to assess and compare designs options. While for all research contexts some performance outcome metric is used (see table 2), researchers should keep in mind, that for game environments measuring performance is usually comparatively easy, as a scoring system is (in most cases) inherently available. For real-world tasks an external (expert) judgment is required, increasing study effort. Self-reported metrics on (agent) performance (commonly used in the included studies) should be used with caution as these might not align with actual performance (e.g., Ashktorab et al., 2020). For virtual research environments an in-depth analysis of the log data might be fruitful (e.g., Tulk et al., 2018).

In the included studies outcomes were measured on team, but not on role or organizational level, which might be due to the novelty of the research field.

10.4.2 Complex problem solving (RQ 2)

In order to understand to what extent CPS is already addressed in the identified studies on human-AI collaboration, an assessment against the CPS taxonomy (Quesada et al., 2005) introduced earlier is performed.

Each of the research contexts identified is evaluated according to the nine characteristics of CPS. Results are aggregated towards research context groups and feature groups (time related, variable related, system behavior related) for clarity. As a result, for each research context group the complexity along the three feature groups is displayed in Table 3. Complexity is assessed and marked as follows:

- *Complex* – marked with ‘X’: assigned to a feature group, if the majority of research contexts fulfilled all three features of complexity
- *Somewhat complex* – marked with ‘(X)’: assigned, if the majority of research contexts fulfilled at least one feature of complexity

Table 3. Research context groups and complexity assessments

Research context group	T	V	SB
Creative content creation		X	(X)
Turn-based cooperative games		(X)	(X)
Real-time cooperative games & simulations	X	(X)	(X)
Design		X	(X)
Negotiation		(X)	X
Investigation	(X)	X	X

T = time related, V = variable related, SB = system behavior related, X = complex; (X) = somewhat complex

In the following, the reasoning for each assessment is explained briefly. The assessments only apply to the research contexts analyzed as part of this literature review without a claim to generalize to other research contexts that could be considered under these labels. Features not mentioned explicitly within the explanations do not fulfill the complexity features for the respective research context groups marked somewhat complex.

Creative content creation. Typically, humans work together with AI systems on the “drawing board” in these research contexts with no time related complexity. The exception among the assessed creative content creation research contexts is the live improvisation of music (McCormack et al., 2019). While drawing, many interrelated variables (e.g., position, shape, coloring, size) need to be considered, making it complex in this regard. With regards to the system behavior, the research contexts are somewhat complex due to the delayed feedback, however, without being opaque or stochastic.

Turn-based cooperative games. The turn-based games within the analysis follow discrete steps without changes to the environment (which are not player induced) and are thus not complex with regards to time. They require users to reason considering a variety of interrelated factors and can thus be considered somewhat complex. Given the adjusted understanding of opacity described earlier, the games analyzed here are not considered opaque as the team can perceive the entire game state. However, due to the delayed feedback they can be considered somewhat complex.

Real-time cooperative games & simulations. With time passing and the environment changing continuously, participants are constantly required to decide, which action to perform. Waiting is also considered an (implicit) action. Thus, real-time cooperative games are considered complex with regards to time. Similar to turn-based games, they also require participants to consider a variety of interconnected variables (somewhat complex). Lastly, besides delayed feedback, some research contexts are stochastic or opaque making them somewhat complex regarding system behavior.

Design. Similar to the creative content generation group, humans create their designs iteratively and the AI provides feedback or suggestions. The research contexts are set in a dedicated design program with no time related complexity. However, a plethora of different, non-linearly related variables need to be considered (e.g., determining if a constructed bridge can carry a certain weight) to successfully achieve the goal, making it complex in this regard. The system behavior can be described as somewhat complex, usually deterministic and transparent but with delayed feedback.

Negotiation. The negotiation is performed in discrete steps (offer, counteroffer, etc.) and is therefore not considered to be complex with relation to time. It requires participants to keep track of several interrelated variables, however, mostly linearly connected, thus considered somewhat complex in this regard. With another AI based system negotiating and the underlying policy of the agent being unknown to the human-AI team, the system behavior can be considered complex.

Investigation. Though time passes continuously and (bad) actors might continue to trade plants during participants deliberating, the system evolves slowly (i.e., low degree of time pressure) and is thus only somewhat time-complex. The team needs to keep track of a large number of interconnected variables and is therefore considered complex in this regard. Lastly, with other people acting within the system, the behavior is opaque, stochastic and provides only delayed feedback (e.g., after offline investigation), making it complex.

In summary, the research contexts are mostly simulations of problems, which are not complex with regards to time as they are usually static, meaning “changes occur only whether participant intervenes” (Quesada et al., 2005). The real-time cooperative games & simulations are an exception. In this research context group, while participants deliberate, the environment can change (“move on”). Many research contexts are not carefully constructed “microworlds” but model real world problems with many, non-trivially interrelated variables and are therefore at least somewhat complex with regards to variables. Similarly, the system behavior is at least somewhat complex for all research contexts due to the delayed feedback, with some research contexts even being stochastic or opaque.

While in AI research reporting on the environment characteristics is common, only few authors explicitly report on the characteristics in the included studies. An exceptions is Song et al. (2021), reporting characteristics in detail and pointing out the before mentioned advantage of game and simulation environments: fine-grained control.

Song et al.'s (2021) study is particularly relevant in the context of this paper, Song et al. (2021) investigate *complexity* as one of their input factors and find significant changes in the AI systems effect on performance: “AI assistance is most beneficial when addressing moderately complex objectives but exhibits a reduced advantage in addressing highly complex objectives”. Only one other study investigates environment changes, similarly reporting changes in strategy (Ong et al., 2012). This sensitivity of results to complexity is well-documented in CPS literature (Frensch & Funke, 1995). We therefore encourage researchers on CPS through human-AI collaboration to investigate the robustness of their findings with regards to complexity, e.g., by using systematic variation of system properties (Frensch & Funke, 1995). This will enable researchers to more precisely describe to which contexts their findings generalize to. While for real-world task practical value might be easier to demonstrate, for research results in games and simulations the argument must be made convincingly, as results on CPS in the laboratory might not necessarily transfer into practice (Frensch & Funke, 1995).

While none of the research contexts are considered complex according to all features they might still be defined as complex as “[o]f all the features studied, none is necessary and sufficient to define a task as ‘complex’” (Quesada et al., 2005). Thus, there is already human-AI collaboration research dealing with (aspects of) CPS that can be used as a foundation to address this recently posed challenge.

10.5 Limitations

Several limitations need to be considered. The search term was informed by the theoretical concepts surrounding the broad term “working together” and focused on “collaboration” and “teaming” due to their similarity and fit to what was described as a research challenge by Dellermann, Ebel, et al. (2019). However, joint work of humans and AI is studied under a variety of labels and even though we additionally considered “cooperation” and “coordination”, potentially not all relevant works are covered. Additionally, this paper builds on a cross-disciplinary literature-based definition of collaboration from a widely cited article in reviewing and excluding search results. Using a different understanding of collaboration might lead to a different set of papers and potentially different results. In combination with the novelty of the research challenge only 22 articles were included. Therefore, we do not claim conclusiveness. On the contrary, our work should be complemented and extended, as the breadth of research contexts in the scientific literature evolves.

Additionally, the conceptualization of “CPS” is, even though having a long research history (Dörner & Funke, 2017), still fuzzy (Quesada et al., 2005). There are many definitions and potential criteria inspired by different understandings of what should be studied under the label of “CPS” (Frensch & Funke, 1995). Here, the formal features of a literature-informed, widely cited taxonomy of CPS tasks was used. In favor of an integrated framework, frequently used criteria such as the necessity to use “creativity as opposed to routine behavior” (Dörner & Funke, 2017) are neglected. Furthermore, some of those formal features of the taxonomy, while widely used for determining complexity (e.g., number of variables Funke, 2012; Quesada et al., 2005) lack clearly defined decision boundaries and are difficult to apply *ex post*, without insight into the research context instantiation process.

10.6 Conclusion and outlook

The nature of work is evolving and is increasingly characterized by complexity, requiring among other things solving complex problems (Jacobs, 2017). Therefore, Dellermann, Ebel, et al. (2019) define CPS through human-AI teams as a major challenge. The goal of this study was to understand to what extent this challenge is already addressed by extant research.

We offer a literature-informed clarification of the two key concepts: human-AI collaboration and CPS. Based on that, we perform a structured literature review to provide an overview regarding the instantiated research contexts used to perform research on CPS through human-AI collaboration. Most research contexts assessed are either creative content generation/ design settings or cooperative games. Applying the features of CPS from literature, most assessed research contexts are not complex with regards to time, but somewhat com-

plex with regards to variables and system behavior. While none fulfill all complexity features, some might still be considered complex. Thus, extant human-AI collaboration literature can be leveraged as a foundation for the new challenge of complex problem solving.

Besides the identified research contexts, we provide insight into the suitability of research context for studying CPS through human-AI collaboration. Games and simulations offer the advantage of flexible adjustments, which is helpful for enforcing characteristics of *collaboration* (e.g., reciprocity) and exploring sensitivity to *complexity* variations (e.g., opacity). Real-world tasks (e.g., drawing) are more difficult to control and results more difficult to measure. However, justification of practical value might be easier for the latter as compared to lab environments (see CPS literature Frensch & Funke, 1995).

We contribute to research in multiple ways. By describing the two key concepts and explicating their defining characteristics, we provide foundational terminology and frame the research challenge more clearly.

For researchers interested in human-AI collaboration through CPS (who are not committed to a specific use case) we provide inspiration by pointing out suitable research context examples. We provide guidance for research context selection or construction by explaining important aspects (e.g., flexibility, measurability) and study design by highlighting the sensitivity to complexity and recommending systematic variation of system properties to achieve more robust results (Frensch & Funke, 1995).

For researchers who have conducted a study we provide a structure along which they can describe their contribution more precisely. Assuming findings generalize better to more similar situations, the taxonomy offers a way to compare research contexts and more precisely describe the relevant class of problems. Given the sensitivity of results to complexity (documented in CPS literature; also shown e.g., by Song et al., 2021), this is important.

Conceptually, going forward we encourage the identification of additional suitable research contexts, e.g., through the evaluation of existing testbeds from reinforcement learning research and proven contexts from CPS literature regarding their suitability for CPS through human-AI-collaboration research. We also advocate to consider additional dimensions of complexity (e.g., technological or organizational) in future work. This should be done with the type of problems in mind that Dellermann, Ebel, et al. (2019) mentioned when posing this research challenge: e.g., managerial, political or science problems – “tasks that so far seem to be at the core of human intellect” (Dellermann, Ebel, et al., 2019).

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

10.7 References

Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M.,

- Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8), 18–28.
- Ashktorab, Z., Liao, Q. V., Dugan, C., Johnson, J., Pan, Q., Zhang, W [Wei], Kumaravel, S., & Campbell, M. (2020). Human-AI Collaboration in a Cooperative Game Setting: Measuring Social Perception and Outcomes. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2). <https://doi.org/10.1145/3415167>
- Bedwell, W. L., Wildman, J. L., DiazGranados, D., Salazar, M., Kramer, W. S., & Salas, E. (2012). Collaboration at work: An integrative multilevel conceptualization. *Human Resource Management Review*, 22(2), 128–145. <https://doi.org/10.1016/j.hrmr.2011.11.007>
- Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special Issue Editorial: Artificial Intelligence in Organizations: Implications for Information Systems Research. *Journal of the Association for Information Systems*, 22(2), Article 10.
- Calderwood, A., Qiu, V., Gero, K. I., & Chilton, L. B. (2020). How Novellists use Generative Language Models: An Exploratory User Study. *IUI '20 Workshops*.
- Davis, N., Hsiao, C., Singh, K. Y., Lin, B., & Magerko, B. (2017). Quantifying Collaboration with a Co-Creative Drawing Agent. *ACM Trans. Interact. Intell. Syst.*, 7(4). <https://doi.org/10.1145/3009981>
- Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., & Ebel, P. (2019). The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In T. Bui (Chair), *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Dörner, D., & Funke, J. (2017). Complex Problem Solving: What It Is and What It Is Not. *Frontiers in Psychology*, 8, 1153. <https://doi.org/10.3389/fpsyg.2017.01153>
- Fischer, A., Greiff, S., & Funke, J. (2012). The Process of Solving Complex Problems. *The Journal of Problem Solving*, 4(1). <https://doi.org/10.7771/1932-6246.1118>
- Frensch, P. A., & Funke, J. (1995). *Complex Problem Solving: The European Perspective*. Taylor and Francis.
- Funke, J. (2012). Complex Problem Solving. In N. M. Seel (Ed.), *Encyclopedia of the Sciences of Learning* (pp. 682–685). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_685
- Gao, X., Gong, R., Zhao, Y [Yizhou], Wang, S., Shu, T., & Zhu, S.-C. (2020). Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks. In *29th IEEE International Conference on Robot and Human Interactive Communication*. Symposium conducted at the meeting of IEEE Robot & Automat

- Soc Robot Soc Japan Korean Robot Soc Furhat Robot EurAi Assoc Italiana Linguistica Computazionale Assoc Italiana Scienze Voce Springer Robotics.
- Geraghty, R., Hale, J., Sen, S., & Kroecker, T. S. (2020). FUN-Agent: A 2020 HUMAINE Competition Entrant. In *MuCAI '20, Proceedings of the 1st International Workshop on Multimodal Conversational AI* (pp. 15–21). ACM. <https://doi.org/10.1145/3423325.3423736>
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, 62, 729–754. <https://doi.org/10.1613/jair.1.11222>
- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Guzdial, M., Liao, N., Chen, J., Chen, S.-Y., Shah, S., Shah, V., Reno, J., Smith, G., & Riedl, M. O. (2019). Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3290605.3300854>
- Jacobs, R. L. (2017). Knowledge Work and Human Resource Development. *Human Resource Development Review*, 16(2), 176–202. <https://doi.org/10.1177/1534484317704293>
- Koch, J., Taffin, N., Beaudouin-Lafon, M., Laine, M., Lucero, A., & Mackay, W. E. (2020). ImageSense: An Intelligent Collaborative Ideation Tool to Support Diverse Human-Computer Partnerships. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1). <https://doi.org/10.1145/3392850>
- Liang, C., Proft, J., Andersen, E., & Knepper, R. A. (2019). Implicit Communication of Actionable Information in Human-AI Teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3290605.3300325>
- Lin, Y., Guo, J., Chen, Y., Yao, C., & Ying, F. (2020). It Is Your Turn: Collaborative Ideation With a Co-Creative Robot through Sketch. In *CHI '20, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM. <https://doi.org/10.1145/3313831.3376258>
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). Ai-Based Digital Assistants. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>
- Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team Effectiveness 1997–2007: A Review of Recent Advancements and a Glimpse Into the Future. *Journal of Management*, 34(3), 410–476. <https://doi.org/10.1177/0149206308316061>

- McCormack, J., Gifford, T., Hutchings, P., Llano Rodriguez, M. T., Yee-King, M., & d'Inverno, M. (2019). In a Silent Way: Communication Between AI and Improvising Musicians Beyond Sound. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. <https://doi.org/10.1145/3290605.3300268>
- McGee, K., Merritt, T., & Ong, C. (2011). What We Have Here is a Failure of Companionship: Communication in Goal-Oriented Team-Mate Games. In *OzCHI '11, Proceedings of the 23rd Australian Computer-Human Interaction Conference* (pp. 198–201). ACM. <https://doi.org/10.1145/2071536.2071568>
- Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>
- Merritt, T., & McGee, K. (2012). Protecting Artificial Team-Mates: More Seems like Less. In *CHI '12, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2793–2802). ACM. <https://doi.org/10.1145/2207676.2208680>
- Ong, C., McGee, K., & Chuah, T. L. (2012). Closing the Human-AI Team-Mate Gap: How Changes to Displayed Information Impact Player Behavior towards Computer Teammates. In *OzCHI '12, Proceedings of the 24th Australian Computer-Human Interaction Conference* (pp. 433–439). ACM. <https://doi.org/10.1145/2414536.2414604>
- Quesada, J., Kintsch, W., & Gomez, E. (2005). Complex problem-solving: a field in search of a definition? *Theoretical Issues in Ergonomics Science*, 6(1), 5–33. <https://doi.org/10.1080/14639220512331311553>
- Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2). <https://doi.org/10.1016/j.im.2019.103174>
- Song, B., Soria Zurita, N. F., Nolte, H., Singh, H., Cagan, J., & McComb, C. (2021). When faced with increasing complexity: The effectiveness of AI assistance for drone design. *Journal of Mechanical Design*, 1–38. <https://doi.org/10.1115/1.4051871>
- Suh, M., Youngblom, E., Terry, M., & Cai, C. J. (2021). AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3411764.3445219>
- Tabrez, A., Agrawal, S., & Hayes, B. (2019). Explanation-Based Reward Coaching to Improve Human Performance via Reinforcement Learning. In *HRI '19: Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 249–257). IEEE Press.
- Tulk, S., Cumings, R., Zafar, T., & Wiese, E. (2018). Better Know Who You Are Starving with: Judging Humanness in a Multiplayer Videogame. In *TechMindSociety '18*,

- Proceedings of the Technology, Mind, and Society*. ACM.
<https://doi.org/10.1145/3183654.3183710>
- van den Bosch, K., Schoonderwoerd, T., Blankendaal, R., & Neerincx, M. (2019). Six Challenges for Human-AI Co-learning. In R. A. Sottolare & J. Schwarz (Eds.), *Lecture Notes in Computer Science: Vol. 11597. Adaptive Instructional Systems* (Vol. 11597, pp. 572–589). Springer International Publishing.
- vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., & Cleven, A. (2009). Reconstructing the Giant: On the Importance of Rigour in Documenting the Literature Search Process. In *17th European Conference on Information Systems (ECIS)*.
https://www.researchgate.net/publication/259440652_Reconstructing_the_Giant_On_the_Importance_of_Rigour_in_Documenting_the_Literature_Search_Process
- Wang, N., Pynadath, D. V., & Hill, S. G. (2016). The Impact of POMDP-Generated Explanations on Trust and Performance in Human-Robot Teams. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (pp. 997–1005). ACM.
- Whitehead, D., Cowell, C. R., Lavorgna, A., & Middleton, S. E. (2021). Countering plant crime online: Cross-disciplinary collaboration in the FloraGuard study. *Forensic Science International: Animals and Environments*, 1, 100007.
<https://doi.org/10.1016/j.fsiae.2021.100007>
- Yanardag, P., Cebrian, M., & Rahwan, I. (2021). Shelley: A Crowd-Sourced Collaborative Horror Writer. In *C&C '21, Creativity and Cognition*. ACM.
<https://doi.org/10.1145/3450741.3465251>
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *CHI '20, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. <https://doi.org/10.1145/3313831.3376301>
- Zhang, C., Yao, C., Liu, J., Zhou, Z., Zhang, W [Weilin], Liu, L., Ying, F., Zhao, Y [Yijun], & Wang, G. (2021). StoryDrawer: A Co-Creative Agent Supporting Children's Storytelling through Collaborative Drawing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.
<https://doi.org/10.1145/3411763.3451785>
- Zhang, G., Raina, A., Cagan, J., & McComb, C. (2021). A cautionary tale about the impact of AI on human design teams. *Design Studies*, 72, 100990.
<https://doi.org/10.1016/j.destud.2021.100990>

11 Publication 2: Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI

Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429

Abstract

Groups of humans or crowds can be remarkable when coming up with ideas. However, not everyone has a group of humans at their disposal to brainstorm. With recent advances in AI, however, generative large language models (LLM) might be capable of contributing ideas in a brainstorming session, turning individual work of a human into joint work of human and AI. It is, however, unclear, how group effects known from human brainstorming groups transfer to such a human-AI setting. In our mixed-method study (qualitative emphasis) with 24 participants, we investigate how a human brainstorms together with the generative LLM ‘GPT-3’, and how they perceived their experience. Our results highlight known effects like cognitive stimulation but also a risk of free riding. We thereby contribute to the understanding of how generative AI, which is becoming broadly available, can be used to address the challenge of human-AI collaboration for solving open-ended problems.

Keywords: Human AI collaboration, Brainstorming, GPT-3, Cognitive stimulation, Free Riding

11.1 Introduction

Solving problems is a key activity for organizations (Nickerson & Zenger, 2004). Brainstorming is a popular technique for groups of humans to develop new ideas on a topic or problem (Hickey et al., 2003; Maaravi et al., 2021; Osborn, 1953). However, not everyone faced with a problem requiring creative ideas has a group of humans cost-effectively available to brainstorm ideas for their problem or is allowed to share the problem with the world. With advances of machines the question arises whether AI can “augment such human problem-solving performance by [...] discovering valuable candidate solutions” (Krogh, 2018) and collaboration of humans and AI systems could be achieved to solve complex problems (Dellermann et al., 2019). The idea to support individual humans on content-level in creative tasks is not new (Wang & Nickerson, 2017), several systems were proposed, e.g., by offering stimuli based on association dictionaries (Althuizen & Reichel, 2016), or by querying social media for inspiration (Siemon et al., 2015). However, such systems did not generate and contribute complete ideas like another human. Traditionally, creativity is described as a human strengths (Dellermann et al., 2019; Krogh, 2018; Lichtenthaler, 2018). With recent advances in generative language models (GLMs), however, human-AI collaboration for creative tasks might be feasible (Gero et al., 2022). Similar to us, Di Fede et al. (2022) suggested exploring using GLMs to support humans in generating ideas.

From group brainstorming literature, several effects are known to influence group performance when generating ideas (Pinsonneault et al., 1999). Examples include group members being stimulated by the exposure to other group members' ideas to come up with additional ideas (*cognitive stimulation*) enhancing performance, or, group members reducing their effort as soon as they are part of group (*free riding*), reducing group performance (Pinsonneault et al., 1999). If a human works together with an AI system that contributes ideas, individual work of a human is transformed into collaborative work of a human and an AI system. While it is known that humans tend to anthropomorphize AI systems (Salles et al., 2020), it is unclear if typical effects known from human brainstorming groups occur in human-AI brainstorming groups too. However, this is important to understand to design effective human-AI collaboration. Therefore, we seek to investigate how individual humans work with and perceive working with such a novel AI system to answer the following research question: *How do individual humans perceive working together with a generative AI system in a brainstorming activity, particularly with regard to known group effects?*

In our exploratory, mixed-method study (qualitative emphasis), participants individually brainstorm with a generative AI and reflect on their experience afterward via a qualitative survey. We show that signs of group effects (e.g., *cognitive stimulation* and *free riding*, Pinsonneault et al., 1999) known from human brainstorming sessions partially occur in such human-AI groups as well. More broadly, we show that an off-the-shelf, non-configured or adjusted GLM (GPT-3) is capable of providing AI suggestions perceived to be helpful and inspirational. We thereby take a first step towards addressing the broader research challenge of designing human-AI collaboration to solve (complex) problems (Akata et al., 2020; Dellermann et al., 2019). More specifically, we explore the team dynamics of humans working with AI systems (Makarius et al., 2020), which could inform designing effective human-AI collaboration (Seeber et al., 2020), with the AI system taking the role of an expert or peer as opposed to a facilitator (Seeber et al., 2020) in an open-ended problem-solving task as opposed to closed-ended decision problem (Krogh, 2018). With powerful GLMs becoming broadly available and embedded into many applications, we believe our research to be timely.

11.2 Background

We briefly highlight *group effects* known from all-human brainstorming groups. We relate our approach to research around supporting brainstorming through (AI) technology and argue that a new type of AI, generative AI might now enable content-level collaboration through its generative capability.

11.2.1 Brainstorming

'Brainstorming' is a creative technique proposed by Osborn (1953) for groups of humans to generate ideas on a creative problem. Osborn (1953) suggested four rules for the group members to follow: (1) delayed judgment, (2) encouragement of wild ideas, (3) focus on

quantity of ideas, and (4) encouragement of improving and combining ideas. Since then, lots of research has been conducted on performing brainstorming effectively. One stream of research investigated how different group settings affected brainstorming performance. Brainstorming could, e.g., be performed face-to-face or electronically (electronic brainstorming, EBS) with verbal or non-verbal expression of ideas, as a group (“real group”) or individually with pooling of ideas (“nominal group”), or with ideas contributed anonymously or non-anonymously (Cooper et al., 1998; Dennis & Williams, 2005; Pinsonneault et al., 1999). Depending on the setting, different effects influence the group’s brainstorming performance. Providing a comprehensive overview of prior research, Pinsonneault et al. (1999) listed 16 effects: six process gains (i.e., performance-enhancing effects) and ten process losses (i.e., performance-reducing effects). We highlight some effects to provide context for our study.

The process gain of *cognitive stimulation*, i.e., the stimulation of new ideas through the utterances of other group members, is only present in *real groups*, as members of nominal groups do not see utterances of other team members (Pinsonneault et al., 1999). Exposure to other’s ideas, e.g., presented via an audio tape (Dugosh et al., 2000) or through a confederate (Paulus et al., 2013) can improve the brainstorming performance, with exposure to more ideas and higher attention paid having a positive effect (Leggett Dugosh & Paulus, 2005). Besides stimulation through others’ ideas, cognitive priming, i.e., subconscious stimulation of the working memory, was also shown to be effective in improving brainstorming performance (Dennis et al., 2013). During verbal brainstorming, the process loss *production blocking* might occur (Diehl & Stroebe, 1987), which refers to group members “being unable to express ideas as they occur”, e.g., because another group member is speaking and they forget their idea (Pinsonneault et al., 1999). One way to reduce or prevent production blocking is by performing EBS, where humans can type their ideas without waiting for others to finish. Even for EBS with *real groups* different settings have been investigated, such as comparing anonymous and non-anonymous idea contribution (e.g., Cooper et al., 1998; Pinsonneault et al., 1999). While the former can prevent the process loss of *Evaluation apprehension*, i.e., “members fear expressing ideas because of potential retaliation” (Pinsonneault et al., 1999), there also is a risk of free-riding, i.e., “members might limit their efforts and contributions by relying on others to accomplish the task” (Pinsonneault et al., 1999), leading to performance loss. These are only some of the effects that were studied to understand performance differences. However, the focus traditionally was on human groups. It is unclear if those process gains and losses also occur in our proposed setting of a human brainstorming with a generative AI system.

11.2.2 Human-AI collaboration in brainstorming

Humans and AI systems working together was discussed under different labels such as human-AI collaboration or hybrid intelligence (Akata et al., 2020; e.g., Dellermann et al., 2019), with the goal to achieve “superior results to those each of them could have accom-

plished separately” (Dellermann et al., 2019). While recent advances in AI enabled machines to outperform humans in several areas, combining humans and AI to solve complex problems might be the next challenge (Dellermann et al., 2019). One aspect discussed in this context was the role of the AI. Bittner et al. (2019) developed a taxonomy for conversational agents, which we, similar to Siemon et al. (2020), interpret more broadly for human-AI collaboration. They differentiate three roles: facilitators “guide users to reach a certain goal or execute a task”, peers “aim to merge into a human group [...] or become a sparring partner for an individual”, and experts “that have certain skills or fields of expertise that differ from those of their human teammates, but mostly act rather reactively upon request”. Concerning ideation or brainstorming, there has been much research on a facilitative level (Wang & Nickerson, 2017). Earlier research on EBS investigated the effects of using different facets of technology to conduct brainstorming sessions. More recently, AI-enabled tools have been used to guide participants through brainstorming sessions, e.g., conversational agents that react to user inputs or record user ideas (Tavanapour et al., 2020). However, such support typically occurs on a meta-level, i.e., the system does not contribute ideas to solve the problem at hand (i.e., process guidance independently of the specific problem). There was less focus on creating AI systems that support brainstorming on content-level. The nature of the task might explain this. Developing a system that implements procedures to guide humans through the brainstorming process is one thing. However, building a system capable of contributing creative ideas like a human might be difficult. Particularly so, *a priori*, it is unknown which brainstorming questions the users will want to use the system for and the system would need to offer suggestions for any question the user might want to brainstorm. While on a facilitative-, meta-level the same procedure might be used across any brainstorming task, suggestions would have to be different for every brainstorming question.

Although it was shown that offering creative stimuli can significantly improve brainstorming performance (cognitive stimulation) only few systems attempt brainstorming content-level support (Siangliulue, Chan, et al., 2015). Althuizen and Reichel’s (2016) prototype used a word associations dictionary from which they pre-selected a subset of words related to their specific problem to offer inspiration. Siemon et al. (2015) developed a prototype to aid humans in brainstorming by showing related social media content. They transformed user’s ideas into search queries for web and social media platforms and showed the queried content generated by others to offer inspiration. While these approaches affect performance processes, they cannot generate (new) ideas in a comparable manner to a human team member. Creativity was traditionally seen as a human strength (Dellermann et al., 2019; Lichtenhaler, 2018), and AI might be considered constrained due to having to draw on historical data (Krogh, 2018). We suggest exploring state-of-the-art GLMs, which might be able to contribute to creative tasks (e.g., Gero et al., 2022; Memmert & Bittner, 2022). With their “Idea Machine” Di Fede et al. (2022) proposed a similar prototype idea to the idea presented in this paper, based on GPT-3, however, they have not (yet) reported empirical data

on its effects. We believe this to be a new approach of identifying useful stimuli, an “important research topic” (Wang & Nickerson, 2017), without having to rely on “building or employing a semantic network” as suggested by Wang and Nickerson (2017).

In supporting a human more holistically on content-level, such a system might be considered to take the role of a *peer* or *expert* (Bittner et al., 2019), or a *creator* in “reflect[ing] creative skills, such as finding many possible solutions or searching for new ideas and developments” (Siemon, 2022).

11.2.3 Content-level support through generative AI

As described above, supporting open-ended, creative problems like brainstorming tasks with AI might be difficult (Krogh, 2018; Lee et al., 2022), as one cannot anticipate all the possible questions to be brainstormed. This makes developing a pool of ideas or suggestions in advance – from which AI system could choose – impossible. Instead, the AI would need to be flexible to address questions as they arise. With traditional natural language processing (NLP) techniques, models were trained for a specific task (e.g., translation, entity recognition). With GPT-3, a generative large language model was developed that outperforms traditional models in many of those tasks but does not require task-specific training (Brown et al., 2020). Even more so, GPT-3 was trained on a broad data set to predict the next word given a certain input. Thus, it can not only perform typical NLP tasks but continue any user input (e.g., Brown et al., 2020; Gero et al., 2022; Lee et al., 2022). Based on this prior research, we argue that such flexible GLMs might allow collaboration with humans in new ways. They neither require pre-selecting a subset of words from a dictionary (Althuizen & Reichel, 2016), limiting practical usability, nor only show social media search results (Siemon et al., 2015). Instead, actual ideas generated on the spot are offered as content-level support. Research on using GLMs is timely as powerful GLMs are becoming available as cloud-hosted services allowing cost-effective usage even for smaller organizations.

In summary, GLMs like GPT-3, an AI-based technology, which might be capable of addressing open-ended problems without requiring to gather training data first, becomes broadly available and embedded into many products. This could enable human-AI collaboration on the open-ended problem of brainstorming, turning individual human work into collaborative work of human and AI. As part of this collaboration the AI would not act as a facilitator but in the role of a peer/expert or creator, contributing ideas similar to a human. However, it is unclear, how group processes affecting performance (e.g., cognitive stimulation, free riding) observed in human groups transfer to this new human-AI setting.

11.3 Method

Our research goal is to gain a rich understanding of the perception of humans brainstorming with our generative AI system. Given the capabilities of state-of-the-art, generative AI systems might open new ways of human-AI collaboration, we argue it might be difficult for participants to imagine how they would (hypothetically) work with such a system. We thus decided to implement a web application to have participants work together with such a system. To gain a rich understanding of the perception of working with the AI, we employed a qualitative approach (Döring & Bortz, 2016, p. 26), more specifically, we conducted a qualitative survey after the brainstorming session. Qualitative surveys are sometimes referred to as “semi-standardized questionnaires” (Döring & Bortz, 2016, p. 403) as only questions are included but no options for answers. Given the novelty of the technology enabling the system to provide ideas similar to a human, this method allowed participants to describe their experience in their own words (Döring & Bortz, 2016, p. 404). Additionally to the subjective experience reported in the survey, we used the system log data of participants working with the web application to get an account of actual usage. This mixed-method-approach (figure 1) allows us to get a more complete picture of the phenomenon (Venkatesh et al., 2013) allowing us to compare different perspectives drawn from different sources of data (Creswell, 2014). We put more emphasis on the qualitative aspect (Creswell, 2014) as our focus is to gain a rich understanding of the perception of working with the AI.



Figure 1. Study method overview

11.3.1 Data collection

11.3.1.1 Environment

For data collection we used a common approach from brainstorming literature (e.g., Siangliulue, Arnold, et al., 2015; Siemon et al., 2015). We set up a brainstorming web application (figure 2) with a task introduction, a timer, the brainstorming question and the text field for users to add their ideas. Submitted ideas were displayed below and could be edited and deleted. Users could request and see AI suggestions next to their own ideas and accepted (i.e., copy) them if they liked. To produce the AI suggestions we used the state-of-the-art, commercial, generative, GLM ‘GPT-3’ at default settings². We did not make any changes to the model including task (brainstorming) or topic-specific (brainstorming regarding sustainability) training. To produce the suggestions we developed a prompt template (table 1), which we populated with up to three randomly sampled participant ideas (i.e., not shared between participants). These ideas act as ‘demonstrations’ to condition the

² The “temperature” parameter was set to 0.9 as recommended by the GPT-3 documentation for “creative applications” (<https://beta.openai.com/docs/api-reference/completions/create#completions/create-temperature>)

GLM’s output (‘few shot learning’, Brown et al., 2020), without requiring model re-training (i.e., adjusting model weights). We included human ideas into the prompt, as one brainstorming rule is to build upon other’s ideas (Osborn, 1953) and the AI could thus “consider” the ideas generated by the human. For developing the template, we employed prompt engineering techniques, particularly itemization (Mishra et al., 2022). We developed a structure consisting of a *task*, the brainstorming *question*, and the start of an answer. To implement the itemization technique we indicated the number of items we expected GPT-3’s output to contain in the task (number of user suggestions + three new suggestions) and used an enumeration to indicate a start of a list (of ideas). We display the post-processed output (removed enumeration, split into individual suggestions).

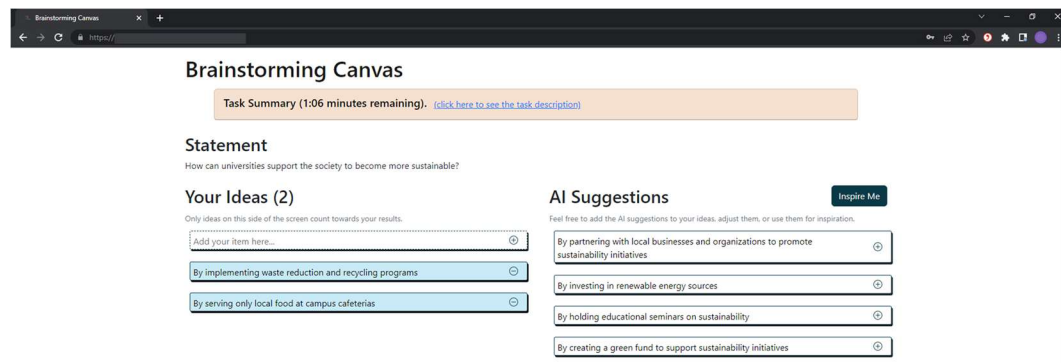


Figure 2. Screenshot of the brainstorming web app (with exemplary ideas)

Table 1. (Empty) prompt template (left) and prompt template populated with three randomly sampled ideas, i.e., ‘demonstrations’, of one user (right)

Prompt template	Prompt template populated with three user ideas
Task: Provide 3 ideas to the question below. Question: How can universities support the society to become more sustainable? Answer: 1. by	Task: Provide 6 ideas to the question below. Question: How can universities support the society to become more sustainable? Answer: 1. By conducting research on sustainability 2. By supporting and implementing sustainable practices on campus 3. By investing in renewable energy sources 4.

According to the described use case users will be aware that suggestions do not originate from other humans, we thus clearly labeled them as AI suggestions to produce a realistic scenario. During the session users need to actively request AI suggestions as it was shown that providing examples on-demand is more effective than automatically providing examples (Siangliulue, Chan, et al., 2015).

11.3.1.2 Participants and procedure

In total 25 participants from three university courses completed the exercise with a completion time of about 25 minutes. We asked users to brainstorm as many ideas as possible (brainstorming rule according to Osborn, 1953) for six minutes. Universities are a frequent topic of brainstorming studies (Baruah & Paulus, 2016; e.g., Dennis & Valacich, 1993; Kohn & Smith, 2011; Paulus et al., 2013), so are societal problems (e.g., Huber et al., 2019; Y. Zhu et al., 2020). We combined the two themes into the question: “How can universities

support the society to become more sustainable?”. It was explained to participants that they could request AI suggestions during the session and could use those suggestions as they liked. As we sought to investigate humans working with the AI system and in brainstorming building ideas upon other’s ideas is encouraged (brainstorming rule according to Osborn, 1953), we asked participants to request AI suggestions at least once. After the session, participants were asked to select their top three ideas (Siangliulue, Arnold, et al., 2015) and fill in the qualitative survey (Döring & Bortz, 2016).

11.3.1.3 Measures

Similar to Siemon et al. (2015) we use two sources of data: recoded actual usage data (i.e., log data) and participants’ reported perception. This triangulation allowed us to get a more in-depth understanding of the usage behavior. For the usage data, we recorded all interactions of the participants with the system with timestamps, including start time of the brainstorming session, submission of every idea, every change to ideas, every suggestion request, and every copying of AI suggestions. For the perceived experience data, we developed a qualitative survey (Döring & Bortz, 2016) consisting of 18, mostly open-ended questions regarding the thought process and experience during the session. We developed the questions based on the group effects to potentially occur as summarized by Pinsonneault et al. (1999) and introduced above.

11.3.2 Data analysis

For the two data sources (i.e., usage data, survey responses) we used different data analysis techniques. For the usage data, we performed descriptive statistical analysis to get a foundational understanding of how the humans worked with the AI system during the brainstorming session. Due to a technical error, some participants were able to submit ideas after the time allocated for the brainstorming session passed. For reporting the results, we filtered out those ideas if not indicated otherwise. To get a more in-depth understanding, we then performed a qualitative content analysis of the open-ended survey responses following Mayring (2014). As our research goal was to understand the group effects within this human-AI setting, we used a deductive approach in which we used the group effects, i.e., the process gains and process losses (e.g., cognitive stimulation, free riding, observational learning) in brainstorming reported in Pinsonneault et al. (1999) as a foundation. We analyzed the responses for indications of such effects and coded them accordingly in MaxQDA (VERBI Software, 2019). We used the descriptions of the group effects by Pinsonneault et al. (1999, Figure 2) as guidance. Thereby we ensured a theoretical grounding with clear reference to our research question as suggested by Mayring (2014). Additionally, we inductively coded the participant’s assessments of the AI suggestions to better understand their perception and what they liked or disliked about the suggestions, and participant’s comments with regard to working with the tool to identify potential improvements for follow-up studies. This additional inductive coding allowed us to gain a more complete understanding of participants’ experiences without “preconceptions of the researcher” (Mayring, 2014, p. 79).

11.4 Results

11.4.1 Overview: working with the AI system

We had 25 study participants completing the exercise in three university seminars on design science research, data-driven solutions for smart cities, and human-centered AI. Due to insufficient completion of the survey one participant was excluded. The remaining 24 participants (female=6, male=18) had an average age of 26.8 years and were bachelor or master students in study programs of the informatics department (except for one biology student), the majority in informatics (n=12) and information systems (n=7). They generated 98 ideas (mean=4.1) during the six-minute brainstorming sessions. As our goal was to investigate the participants' perception of working with the AI systems, we asked them to request AI suggestions at least once. However, most participants made multiple requests for AI suggestions (mean=4.5). Participants made 109 requests, received 326 AI suggestions, and accepted (i.e., copied) 156 AI suggestions (mean=6.5). An overview regarding the timelines for each brainstorming session is shown in figure 3. On average, participants added 1.1 ideas prior to making their first request 56.8 seconds into the session. Across participants, splitting the session in 1-minute intervals, participants jointly made 19, 19, 19, 16, 14, and 22 requests respectively. Examples of AI suggestions include (original spelling): “By increasing the number of vegan/vegetarian options in cafeterias”, “By Give awards to students/staff/teams who have contributed to making the university more sustainable”, and “By have a "sustainability week" where the school shuts down all unnecessary lights and appliances”.

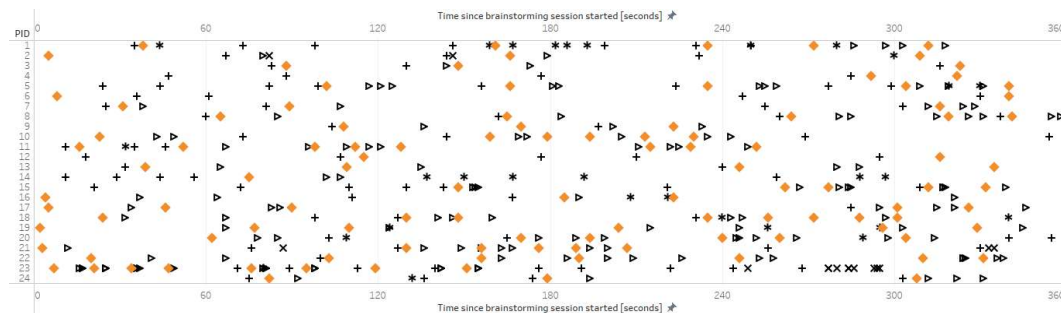


Figure 3. Timeline for brainstorming sessions (+ added idea, * adjusted node, ♦ requested AI suggestions, ▷ accepted/copied AI suggestion)

11.4.2 Cognitive stimulation/synergy

Cognitive stimulation refers to “utterance[s] of members may contain task related stimuli that elicit new ideas from other members” (Pinsonneault et al., 1999), leading to a performance gain. For the reason discussed above participants had to actively request suggestions, which might act as stimuli. While we asked participants to make at least one request for AI suggestions (to allow the investigation of the perception of working with an AI system), all but one participant performed multiple requests (min=1, max=8, mean=4.5). When asked for their reasoning as to why they *requested* AI suggestions, multiple participants stated that they were curious and expected to receive ideas or inspiration, or that they

ran out of ideas (e.g., “I ran out of ideas or for the moment didn't think of any that quickly and then got ideas or inspiration from them” (P05)). When asked about the AI suggestions they *received*, participants explained that they did feel inspired through the suggestions. Participants highlighted that the suggestions helped them to explore new areas of ideas:

- The AI “inspired me to think in different directions” and “I really like it. It helps to broaden the horizon and get inspired to think in differs ways/areas than before. After I took the suggestions I was able to come up with a few additional ideas I didn’t think of before. And that is just because the AI inspired me to explore this particular area.” (P15)
- AI suggestions “hint[ed] me in some direction of thinking.” (P14)
- The AI “gave inspiration to think into different directions (e.g.: in the beginning I mainly thought about ideas on how to educate the public, the AI gave me the idea to come up with more direct ideas/things to do)” (P05)

An example provided by one of the participants (P05) was that the AI suggested an idea regarding using solar panels and the participant then “thought about gaining energy from wind”. The responses show that participants expected to receive new ideas and actually felt to have gained new ideas from the AI. We therefore conclude that the AI system might lead to the process gain of cognitive stimulation.

11.4.3 Free riding

The process loss free riding is the “motivated, intentional withdrawal of efforts”, where group “members might limit their efforts and contributions by relying on others to accomplish the task” (Pinsonneault et al., 1999). Some participants acknowledged having relied on the AI stating that working with the AI is “easier than thinking alone” (P09) or working without it “would have been too much thinking” (P18).

From literature, several underlying factors are known that lead to free riding including (1) perceived dispensability of one’s effort and (2) diffused responsibility (Pinsonneault et al., 1999). With regard to (1) perceived dispensability, several participants stated that they do not feel to be knowledgeable in the topic and therefore welcomed the support by an AI system (P13). They also felt that the brainstorming task required creativity and they find themselves not to be a “creative person” (P16). These are examples where people might feel dispensable and might therefore tend to free ride.

With regard to (2) diffused responsibility, we asked participants “To what degree do you feel responsible for the results, i.e., final list of ideas (0-100%)?” and asked them to justify their response. We report the results on the responsibility scores (figure 4) separately for five participants (group 2), as these participants reported they did not feel to be able to provide an estimation (P12), provided contradictory information (P21), or provided answers more generally related to sustainability (P1, P2, P8). For the remaining participants, most participants (14) reported a higher felt responsibility (mean=64.7%) as compared to

the share of ideas they contributed (mean=41.0%), for two participant the values were aligned, and three participants felt a lower responsibility score as compared to their contribution.



Figure 4. Reported felt responsibility for the final list of results (red squares) vs. share of final ideas by the human (blue circles)

Table 2. Participants’ justification and explanation for the reported responsibility score (RS), and actual share of contribution towards the final list of results (AS)

PID	Justification	Explanation	RS	AS
P11	Share of ideas	“I only wrote two or three ideas by myself and accepted a lot more suggestions from AI.”	20%	15.8%
P03	Share of ideas	“The final list consists of half of my own ideas.”	50%	60.0%
P16	Share of top ideas	“My top three choices are two of my own and one AI generated”	66%	42.9%
P07	Curation	“I typed 5 answers from myself and got 5 from AI suggestion system. So I can say 51% because I picked 5 of AI so it was still in my responsibility.”	51%	44.4%
P22	Curation	“I chose what to select and what not to.”	95%	6.7%

Many participant justified their responsibility score with the share of ideas they contributed (see table 2, e.g., P11, P03), while one participant referred to the distribution among the top three choices (P16). Several participants explain to feel high(er) responsibility for the result as they reviewed the suggestions and made the final selection, i.e., curated the results (e.g., P7, P22). Thus, different justifications for the reported responsibility scores exist, with some participants finding it difficult to provide a justification altogether, surfacing the need for further exploratory research perceived responsibility formation.

Consistent with prior findings, the results show participants on average claiming higher responsibility as compared to the share of ideas contributed. Nonetheless, except for two all participants claim lower than full responsibility for the final results, which might hint – again – at a known factor for free riding.

To “free ride” during the task, participants would need to generate suggestions and accept (i.e., copy) them. We thus analyzed the log data: in total 156 suggestions were accepted, only three of those suggestions copied were adjusted by the participants. However, participant’s survey responses do not necessarily hint at a withdrawal of efforts, as participants seem to have engaged with the suggestions. Several participants praised the idea quality, as “well thought out suggestions” (P09), which are “perfectly understandable” (P16), “diverse” (P03, P11) and were presented “in the right amount” (P03). They used similar arguments when explaining when and why they accepted suggestions, describing them as “smart” (P17), “good ideas” (P02), “short, precise and I could agree to their message”

(P24)”. This aligns with their explanations as to why they *accepted* many suggestions *without making changes*:

- “I did not make changes because AI suggestions were good.” (P11)
- “Some of them covered blindspots, like composting and rainwater collection” (P20)
- I “did not make changes: because they were clear, and it sometimes didn't make sense to make an idea more specific/less broad” (P05)
- “I did not adjust when they were straightforward okay” (P20)

This high level of perceived quality of suggestions is also reflected at the stage of idea selection. After completing the brainstorming exercise, participants were presented with the final list of their ideas including AI suggestions they had accepted (i.e., copied). Participants were asked to “select a diverse set of your top 3 ideas”. While the majority of ideas selected as best originated from a human, 35.2% percent originated from the AI system, with 17 participants (70.8%) selecting at least one accepted AI suggestion as one of their best idea (figure 5). Those final ideas based on AI suggestions included, e.g.,

- “By Introduction of new study programs with a focus on sustainability” (P01)
- “By Establishing a "Green Fund" to support student-led sustainability initiatives” (P21)
- “By implementing waste reduction and recycling programs” (P17)
- “By collecting rain water for reuse in toilet flushing or landscape irrigation” (P20).

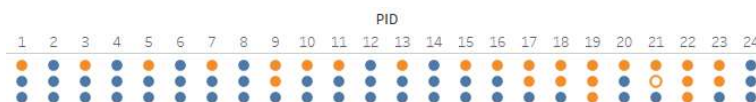


Figure 5. Participants’ selected top 3 ideas by origin: ai=orange, human=blue (for P21 one selected idea was based on an AI suggestion after the brainstorming time had passed)

We thus conclude that the high number of accepted, non-adjusted suggestions might at least partially be related to the high perceived quality of AI suggestions and thus might not necessarily indicate free riding. The fact that many of those suggestions were not adjusted by the participants might be partially attributed to the time constraint, and to the fact that, due to the interface design, several participants were not aware adjustments were possible, even though it was explained during the introductions (e.g., P18).

This active engagement of the participants with the suggestions is also reflected in the participants’ reasoning as to why they decided not to copy AI suggestions, most notably due to a lack of formal- and content-related quality of the ideas, with grammar errors or ideas perceived to be “basic”, “blunt” (P04), or not realistic (P16). With regard to diversity of suggestions, there was disagreement among participants with some praising the diversity

(e.g., P03) and others stating a perceived lack of diversity (e.g., P22). On a similar note, there was disagreement with regard to the relative distance of suggestions in relation to the human entered ideas, with some participants stating suggestions were not based on the ideas (e.g., P02) and others stating the suggestions were basically their ideas phrased in different words (e.g., P10).

To summarize, free riding, i.e., the intentional withdrawal of efforts, occurs e.g., when people (1) feel their skills are dispensable, or (2) when responsibility is unclear (Pinsonneault et al., 1999). We see three indications for free riding, with some participants pointing out they did feel to lack qualification for the task (creativity, domain knowledge), most participants' reporting not feeling fully responsible for the results, and participants accepting many of the AI suggestions without making changes. However, the analysis of the qualitative survey responses shows a different picture: participants did critically engage with the suggestions, highlighting both positive and negative aspects around the suggestions and explaining their interaction behavior. According to those responses, the high number of copied suggestions might be due to the good quality of those suggestions. This is particularly underlined by the fact that many of the accepted suggestions were select into their final list of ideas by the participants.

11.4.4 Observational learning and negative productivity matching

Previous research shows that people within group brainstorming adjust their effort according to the performance of other group members. This can lead to performance gains when members learn from others and try to “imitate best performers” (“observational learning”) (Pinsonneault et al., 1999), or lead to performance losses, when group members adjust their “productivity to a baseline level” (“Negative productivity matching”) (Pinsonneault et al., 1999). Multiple participants reported to have requested AI suggestions because they wanted to understand what is expected of them by seeing what the AI system would suggest, e.g., to “check if my thinking and ideas were going into the right direction” (P10), to “check if my answers fulfill the requirements of the task” (P15), or “I never thought about this topic and wanted to understand the direction” (P17). We assume that this partly has to be attributed to the study setting, where the participants thought there to be a “correct” way of writing ideas and assumed this would be reflected in the implementation of the AI. However, this might indicate participants calibrating themselves with regard to the AI performance. Thus, providing AI suggestions needs to be done with caution, particularly in the early stages of the brainstorming session if those suggestions implicitly set the baseline for the perceived performance expectations of the human. On the other hand, if humans indeed adjust to the AI, this might open an opportunity to shape humans' performance by careful design.

11.4.5 Cognitive inertia and cognitive interference

Cognitive inertia refers to the performance reducing effect when individuals “embark on a single train of thought, which limits creativity and productivity” (Pinsonneault et al., 1999).

Such an effect occurs during individual brainstorming (nominal groups) as individuals do not receive external stimuli. In our study, participants did not brainstorm with other humans, but they also did not brainstorm alone (but with the AI system). Many participants felt inspiration through exposure to the AI suggestions (cognitive stimulation). However, when asked how they thought the suggestions influenced them some reported that while the suggestions “gave [them] a direction to think about”, “they also closed other directions down” (P18) and that the suggestions “made [them] come up with ideas that were close to the ones the AI provided (P05)”, with some suggestions being “parallel to [their] ideas” (P07). As mentioned before, some participants felt that the AI re-stated ideas. One participant even explained: “I prefer to first brainstorm alone, and then, only when I’m “stuck” to look at the AI suggestions, because I’d otherwise feel like the ideas of the AI would bias me and prevent myself from being/staying creative” (P05).

On a technical level, the perception of AI suggestions being similar might be explained by the AI system being conditioned by the human input therefore making suggestions related to the input. On this technical level, this could be prevented as part of improved prompt engineering, e.g., by providing a diverse set of sample ideas to condition the AI to “think in different directions”. On a more abstract level, the phenomenon of a narrowing of creativity due to stimuli is known as *fixation* (Sio et al., 2015). We thus suggest to employ existing literature on fixation for developing such socio-technical systems.

11.4.6 Group setting preference: alone, with AI, or with another human

Given their experience, we asked participants if they would have rather worked alone as compared to with the AI, and, if they would rather have worked with another human or with the AI. Most participants preferred working with the AI as compared to alone (79%), stating that “it is nice to have some assistance to be more productive (P07)” and feeling that working with the AI “was easier and faster” (P21) and they were able to “gather[...] many ideas” (P10). However, the AI system does not replace another human, with many participants stating that they would have preferred to work with another human as compared to working with the AI, e.g., due to the “social” aspect (P12) or because they believed it could be more fun (P05, P10). On the other hand, several participants stated they would not have preferred to work with another human, as this would result in “too much discussion” (P18) “tak[ing] up more time” (P11), whereas working with the AI “saves your time” (P01). Several participants stated that they do see working with the AI system and another human as mutually exclusive (P23), but instead suggested a mixture (P03), e.g., the AI systems could be used in a “preliminary step that can lead to a discussion with a human” (P24) or “Having another person to discuss AI suggestions would be nice” (P23).

Participant’s preferences, i.e., rather work with an AI than alone, but rather work with another human than with the AI (exclusively), fit the situation explained in the introduction: the AI might improve (perceived) brainstorming performance for individuals but cannot (yet) replace a human teammate.

11.5 Discussion and future research

11.5.1 Process effects in human-AI groups

In our study, we propose a human to brainstorm with a generative AI. The setting is similar to all human brainstorming sessions, in that group members contribute as many ideas as possible and build onto other group members' ideas; but different, in that the human requests AI suggests (as this was shown to be most effective, Siangliulue, Chan, et al., 2015). We show that certain group effects known to occur within all-human settings (Pinsonneault et al., 1999) also seem to occur in a human-AI setting. Particularly, we saw the potential for the process gain '*cognitive stimulation*', with participants reporting that they felt they were faster in gathering ideas and covered more aspects as compared to working alone. This is underlined by the large portion of participants wanting to work with the AI instead of working alone. Such a setup could therefore materialize the ambition of increased performance through human-AI collaboration as compared to humans working alone as described by Dellermann et al. (2019). Particularly so, as besides the positive group effect there is a direct effect of the AI system adding ideas. Indeed, participants not only copied many AI suggestions but even selected some of them into their best three ideas, indicating a potential performance improvement. Future research should investigate the qualitative differences between these ideas to improve understanding of human and AI creativity.

However, we also see the risk of the process loss of '*free riding*' to occur. There have been calls for research on free riding in human-AI collaboration scenarios (Siemon & Wank, 2021). Our results are certainly less clear as compared to what Siemon et al. (2015) reported for their study with an AI-like support system, stating no-free riding occurred when using their prototype. A potential explanation could be that the quality of stimuli improved with advances of AI, increasing the participants' perceived dispensability of efforts, a known factor leading to free riding. Another explanation could be the difference in scenario or difference in measurement, with us discussing participants' explanations of their behavior and Siemon et al. (2015) asking participants directly, if they had "exert[ed] less effort" knowing they were supported and their "effort is not instrumental obtained to the outcome". Occurrence of free riding would certainly be consistent with previous research on humans collaborating with virtual assistants (Stieglitz et al., 2022) or on AI assisted decision-making on closed-ended problems, where humans sometimes show signs of "overreliance", accepting AI suggestions even though they are incorrect (Buçinca et al., 2021). This might similarly occur in this brainstorming scenario, particularly given that AI suggestions are not necessarily truthful (Lin et al., 2022). We suggest for future research to investigate how free riding and "withdrawal of efforts" should be conceptualized and measured in these human-AI settings. Humans contributing less ideas when being part of a group is a sign of "withdrawal of efforts" and thereby free riding in traditional all-human real groups. However, in human-AI groups, such a narrow view on the quantity of human ideas might – as our study indicates – draw an incorrect picture, as humans might be engaged in curating AI

suggestions instead of developing ideas. A more nuanced explanation might be humans feeling they can reduce their effort to “maintain cognitive resources and enhance efficiency in work” (Stieglitz et al., 2022, p. 758) believing the AI to compensate for them, described by Stieglitz et al. (2022) as “smart loafing”. This new role might have contributed to some participants finding it difficult to explain their degree of responsibility for the results.

Overall, we observed signs for both performance enhancing and performance reducing effects. We thus believe it will be important for future research to quantify those performance differences to see the net effect. Typically, brainstorming performance of groups is assessed by measuring quantity and quality of ideas, where quantity is the number of non-redundant ideas (e.g., Michinov, 2012; Ritter & Mostert, 2018) and quality might be assessed, e.g., with regard to usefulness, originality, diversity etc. (Ritter & Mostert, 2018; e.g., Siangliulue, Arnold, et al., 2015). Measuring performance quantitatively, however, might be unfit here, as the AI can generate an unparalleled number of ideas. Thus, we suggest to adjust performance measurement. As explained earlier with regard to study setup, typically, nominal groups are compared to “real” groups. We, however, argue both from the research path on human-AI collaboration/hybrid intelligence as well as from a practical perspective, ultimately the comparison between (1) human (or another appropriate baseline such as human + search engine or human + ostensible ideas from a social other) and (2) human + AI will be vital. From the research perspective on hybrid intelligence as set out by Dellermann et al. (2019, p. 640) this comparison is important to understand whether the human-AI collaboration actually achieves the goal of “superior results”, i.e., if the “socio-technical system achieves a performance in a specific task that none of the involved agents, whether they are human or artificial, could have achieved without the other”. Having the human work individually might be a reasonable baseline from a practical perspective, as humans might be faced to either perform the task alone or with the AI, but might not have the option to work with other humans.

Our study offers insights into designing human-AI collaboration for brainstorming and thereby practical implications. We show that generative AI (e.g., GPT-3) may be used to generate helpful suggestions in brainstorming settings. However, such suggestions need to be designed carefully, as they might implicitly be perceived as a performance expectation baseline by the humans. We therefore suggest to make use of prompt engineering to improve suggestion quality both on formal (e.g., grammar) and content (e.g., diversity) level. The latter is particularly important as some participants stated to have found the suggestions to lack diversity, which could lead to a phenomenon called fixation (Sio et al., 2015). Even with those improvements, however, to reduce the human relying on the AI, we suggest to make transparent that the system does not have additional task information. We also suggest careful interaction design to prevent the human uncritically accepting AI suggestions and being invited to free ride. There is much research available on increasing engagement and calibrating trust from AI-assisted decision making (e.g., Bućinca et al., 2021), e.g., through timing or partial disclosure of information.

11.5.2 Generative AI for collaboration on content-level when solving problems

Generating ideas is an important part of solving open-ended problems. Lots of research was conducted on facilitating brainstorming, e.g., by offering process guidance through technology, more recently with AI enablement. However, whereas meta-level facilitation might be conducted independently of the question, content-level collaboration requires engagement with the specific question at hand. We show that generative AI can enable such collaboration. Due to its flexible design there is no need to anticipate the brainstorming questions or gather training data in advance, as suggestions are generated on the spot.

There are some limitations to be mentioned. We only used one brainstorming question to ensure comparability between participants. While this is a limitation, we want to stress that for our study we did not make any changes to the generative AI (GPT-3) we used. More specifically, we did neither gather training data nor did we train the model. On the contrary, the brainstorming question could simply be replaced in the prompt template string (table 1) to produce different results. We therefore argue that our approach is re-usable and our results might be generalizable to other brainstorming questions, which are sufficiently general and can be answered with public knowledge. That we only used one model (GPT-3) might be considered another limitation. We decided to use GPT-3 as it was shown to have a high performance across many tasks (Brown et al., 2020) and is easily usable via API. Our approach, however, is flexible to be used with other models. Lastly, while we used prompt engineering techniques for prompt construction, results potentially can be improved with more elaborate prompt engineering, potentially even including examples of ideal ideas in brainstorming (Q. Zhu & Luo, 2022).

With our initial study, we offer a transparent, re-usable approach as a foundation. We show that GLMs like GPT-3 have the potential to enable new forms of collaboration for open-ended problems (as opposed to repetitive, closed-ended decision tasks) which are important in organizations (Nickerson & Zenger, 2004), potentially enabling AI to contribute like an expert/a peer on content-level instead of supporting as a facilitator on meta-level. This might open up the opportunity to support humans in a creative activity, which seemed to be one of the core competencies of humans (Dellermann et al., 2019; Lichtenthaler, 2018). We thereby work towards the research challenge of solving complex problems with human-AI collaboration (Akata et al., 2020; Dellermann et al., 2019; Krogh, 2018). More specifically, Makarius et al. (2020) called for research on “team dynamics of working side-by-side with AI systems”, Dellermann et al. (2019) described “superior performance” as a core characteristic of hybrid intelligence, and Seeber et al. (2020) pose the question of how to design machine teammates for effective collaboration. With our study we work towards addressing those challenges in exploring the use of an established lens from all-human brainstorming groups in this new human-AI setting, to better understand team dynamics,

particularly performance-affecting processes and ultimately to inform the design of generative AI-based systems to work effectively with humans. While already with this minimal approach most participants preferred working with the system as compared to working alone, we suggest to build on the approach to further improve the collaboration. Besides investigating the robustness of our results, particularly with regard to the limitations (brainstorming question, model, prompt), we propose future research should investigate the integration into settings with multiple humans, existing facilitation approaches on meta-level, and different idea generation techniques.

More broadly, research on this setting, particularly due to its open-ended nature, could inform the discussion around augmentation vs. automation through AI (e.g., Raisch & Krakowski, 2020) and AI-based systems team roles (Bittner et al., 2019; e.g., Siemon, 2022). While cognitive stimulation through the AI resulting in humans developing new ideas might be considered augmentation, the AI adding ideas without human adjustments might be automation. A better understanding of the relative strength of the performance affecting processes (gains and losses) will help to strike a balance to achieve high performance and potentially *superior performance* (Dellermann et al., 2019). Within our setting, human participants took both a contributor (own ideas) and a curator role (selecting from AI ideas). The AI system contributing ideas and thereby working on content-level can be considered a *creator* according to Siemon (2022) team roles. It combines characteristics of both the peer and the expert role according to Bittner et al.'s (2019) team roles. Future research should investigate managing attention (higher attention to others' ideas leads to more ideas generated, Leggett Dugosh & Paulus, 2005) and how this human role change might open up new opportunities for enabling less skilled humans (Rafner et al., 2021), and for reconfiguring (Afiouni & Pinsonneault, 2022) or repurposing jobs (Seeber et al., 2020).

11.6 Conclusion

With advances of generative, large language models new ways of human-AI collaborations might become feasible. In our exploratory mixed-method study, we investigate humans' perception of working with a generative AI in a brainstorming setting. We show that certain effects known from all-human brainstorming groups seem to appear within human-AI groups too, particularly cognitive stimulation and free riding. We discuss implications for designing such sociotechnical systems and for advancing on the challenge of human-AI collaboration for complex problem solving (Dellermann et al., 2019).

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

11.7 References

- Afiouni, R., & Pinsonneault, A. (2022). Ripples of Change – An AI Job Crafting Model for Human-in-Control. In *Proceedings of the 55th Annual Hawaii International Conference on System Sciences*. <https://hdl.handle.net/10125/79969>
- Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8), 18–28.
- Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of Management Information Systems*, 33(1), 11–44. <https://doi.org/10.1080/07421222.2016.1172439>
- Baruah, J., & Paulus, P. B. (2016). The Role of Time and Category Relatedness in Electronic Brainstorming. *Small Group Research*, 47(3), 333–342. <https://doi.org/10.1177/1046496416642296>
- Bittner, E., Oeste-Reiß, S., & Leimeister, J. M. (2019). Where is the Bot in our Team? Toward a Taxonomy of Design Option Combinations for Conversational Agents in Collaborative Work. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. <https://doi.org/10.24251/HICSS.2019.035>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Cooper, W. H., Gallupe, R. B., Pollard, S., & Cadsby, J. (1998). Some Liberating Effects of Anonymous Electronic Brainstorming. *Small Group Research*, 29(2), 147–178. <https://doi.org/10.1177/1046496498292001>
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4. ed., international student ed.). Sage.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>

- Dennis, A. R., Minas, R. K., & Bhagwatwar, A. P. (2013). Sparking Creativity: Improving Electronic Brainstorming with Individual Cognitive Priming. *Journal of Management Information Systems*, 29(4), 195–216. <https://doi.org/10.2753/MIS0742-1222290407>
- Dennis, A. R., & Valacich, J. S. (1993). Computer brainstorms: More heads are better than one. *Journal of Applied Psychology*, 78(4), 531–537. <https://doi.org/10.1037/0021-9010.78.4.531>
- Dennis, A. R., & Williams, M. L. (2005). A Meta-Analysis of Group Side Effects in Electronic Brainstorming: More Heads are Better than One. In *International Journal of e-Collaboration*. <https://doi.org/10.4018/JEC.2005010102>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Pers. Soc. Psychol.*, 53(3), 497–509. <https://doi.org/10.1037/0022-3514.53.3.497>
- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-41089-5>
- Dugosh, K. L., Paulus, P. B [P. B.], Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *J. Pers. Soc. Psychol.*, 79(5), 722–735. <https://doi.org/10.1037/0022-3514.79.5.722>
- Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing using Language Models. In F. ` Mueller, S. Greuter, R. A. Khot, P. Sweetser, & M. Obrist (Eds.), *Designing Interactive Systems Conference* (pp. 1002–1019). ACM. <https://doi.org/10.1145/3532106.3533533>
- Hickey, A. M., Davis, A. M., & Kaiser, D. (2003). Requirements Elicitation Techniques: Analyzing the Gap between Technology Availability and Technology Use. *Comparative Technology Transfer and Society*, 1(3), 279–302. <https://doi.org/10.1353/ctt.2003.0026>
- Huber, B., Shieber, S., & Gajos, K. Z. (2019). Automatically Analyzing Brainstorming Language Behavior with Meeter. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–17.
- Kohn, N. W., & Smith, S. M. (2011). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 25(3), 359–371. <https://doi.org/10.1002/acp.1699>
- Krogh, G. von (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404–409. <https://doi.org/10.5465/amd.2018.0084>

- Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI'22: Conference on Human Factors in Computing Systems*.
- Leggett Dugosh, K., & Paulus, P. B. (2005). Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology*, 41(3), 313–320. <https://doi.org/10.1016/j.jesp.2004.05.009>
- Lichtenthaler, U. (2018). Substitute or Synthesis: The Interplay between Human and Artificial Intelligence. *Research-Technology Management*, 61(5), 12–14. <https://doi.org/10.1080/08956308.2018.1495962>
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Maaravi, Y., Heller, B., Shoham, Y., Mohar, S., & Deutsch, B. (2021). Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science*, 15(6), 1431–1464. <https://doi.org/10.1007/s11846-020-00400-5>
- Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. <https://doi.org/10.1016/j.jbusres.2020.07.045>
- Mayring, P. (2014). *Qualitative Content Analysis - Theoretical Foundation, Basic Procedures and Software Solution*.
- Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
- Michinov, N. (2012). Is Electronic Brainstorming or Brainwriting the Best Way to Improve Creative Performance in Groups? An Overlooked Comparison of Two Idea-Generation Techniques. *Journal of Applied Social Psychology*, 42, E222-E243. <https://doi.org/10.1111/j.1559-1816.2012.01024.x>
- Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 589–612).
- Nickerson, J. A., & Zenger, T. R. (2004). A Knowledge-Based Theory of the Firm—The Problem-Solving Perspective. *Organization Science*, 15(6), 617–632. <https://doi.org/10.1287/orsc.1040.0093>

- Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.
- Paulus, P., Kohn, N. W., Arditti, L. E., & Korde, R. M. (2013). Understanding the Group Size Effect in Electronic Brainstorming. *Small Group Research*, 44(3), 332–352. <https://doi.org/10.1177/1046496413479674>
- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., & Sherson, J. (2021). Deskill, Upskill, and Reskill: a Case for Hybrid Intelligence. *Morals & Machines*, 1(2), 24–39. <https://doi.org/10.5771/2747-5174-2021-2-24>
- Raisch, S., & Krakowski, S. (2020). Artificial Intelligence and Management: The Automation-Augmentation Paradox. *The Academy of Management Review*. Advance online publication. <https://doi.org/10.5465/2018.0072>
- Ritter, S. M., & Mostert, N. M. (2018). How to facilitate a brainstorming session: The effect of idea generation techniques and of group brainstorm after individual brainstorm. *Creative Industries Journal*, 11(3), 263–277. <https://doi.org/10.1080/17510694.2018.1523662>
- Salles, A., Evers, K., & Farisco, M. (2020). Anthropomorphism in AI. *AJOB Neuroscience*, 11(2), 88–95. <https://doi.org/10.1080/21507740.2020.1740350>
- Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2). <https://doi.org/10.1016/j.im.2019.103174>
- Siangliulue, P., Arnold, K. C., Gajos, K. Z., & Dow, S. P. (2015). Toward Collaborative Ideation at Scale. In *CSCW '15 Proceedings* (pp. 937–945). ACM. <https://doi.org/10.1145/2675133.2675239>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Siemon, D. (2022). Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation*, 31(5), 871–912. <https://doi.org/10.1007/s10726-022-09792-z>
- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.
- Siemon, D., Li, R., & Robra-Bissantz, S. (2020). Towards a Model of Team Roles in Human-Machine Collaboration. In *ICIS 2020 Proceedings*. https://aisel.aisnet.org/icis2020/is_workplace_fow/is_workplace_fow/7

- Siemon, D., & Wank, F. (2021). Collaboration With AI-Based Teammates – Evaluation of the Social Loafing Effect. In *PACIS 2021 Proceedings*. <https://aisel.aisnet.org/pacis2021/146/>
- Sio, U. N., Kotovsky, K., & Cagan, J. (2015). Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies*, 39, 70–99. <https://doi.org/10.1016/j.destud.2015.04.004>
- Stieglitz, S., Mirbabaie, M., Möllmann, N. R. J., & Rzycki, J. (2022). Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. *Information Systems Frontiers*, 24(3), 745–770. <https://doi.org/10.1007/s10796-021-10201-0>
- Tavanapour, N., Theodorakopoulos, D., & Bittner, E. A. C. (2020). A Conversational Agent as Facilitator: Guiding Groups Through Collaboration Processes. In P. Zaphiris & A. Ioannou (Eds.), *Lecture Notes in Computer Science. Learning and Collaboration Technologies. Human and Technology Ecosystems* (Vol. 12206, pp. 108–129). Springer International Publishing. https://doi.org/10.1007/978-3-030-50506-6_9
- Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the Qualitative–Quantitative Divide: Guidelines for Conducting Mixed Methods Research in Information Systems. *MIS Quarterly*, 37(1), 21–54. <https://doi.org/10.25300/MISQ/2013/37.1.02>
- VERBI Software. (2019). *MAXQDA 2020* [Computer software].
- Wang, K., & Nickerson, J. V. (2017). A literature review on individual creativity support systems. *Computers in Human Behavior*, 74, 139–151. <https://doi.org/10.1016/j.chb.2017.04.035>
- Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, 2, 1825–1834. <https://doi.org/10.1017/pds.2022.185>
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2020). Creativity: Intrapersonal and Interpersonal Selection of Creative Ideas. *The Journal of Creative Behavior*, 54(3), 626–635. <https://doi.org/10.1002/jocb.397>

12 Publication 3: The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3

Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/107289>

Abstract

Generative language models (GLM) like GPT-3 can support humans in creative tasks. Such systems are capable of generating free-text output based on a provided input prompt. Given the outputs' sensitivity to the prompt, many techniques for prompt engineering were proposed both anecdotally in social media and increasingly in literature. It is, however, unclear if and how such a system and such techniques can be employed in creative contexts such as for generating ideas. In our study, we investigate the effects of using six prompt engineering techniques. For each combination of techniques, we have GPT-3 generate ideas for an exemplary scenario. The ideas are rated according to novelty and value. We report on the effects of the (combinations of) prompt engineering techniques. With our study, we contribute to the emerging field of prompt engineering and shed light on supporting idea generation with GLMs, showing a pathway to embedded GLM capabilities.

Keywords: brainstorming, generative language model, GPT-3, human-AI collaboration, prompt engineering

12.1 Introduction

Generative language models (GLM) like GPT-3 (Brown et al., 2020) have gained increasing attention for their ability to support humans in various creative tasks. Such systems are capable of generating free-text output based on a provided free-text input, typically referred to as a *prompt*. With recent advances, users can even directly interact with such GLMs via chat (e.g., ChatGPT). While such an interaction directly with the model is possible, GLM capabilities can be embedded into products, essentially abstracting away the complexities of interacting with the GLM for the users. An example of such a system is depicted in Figure 1. In this idea generation application prototype, the user can add a question, add first ideas, and request AI ideas if desired. The user can review the suggested ideas and select or ignore them as they wish. The suggestions are generated dynamically in the background. The user does not need any knowledge on how to interact with GLM but can still benefit from its capabilities. This removes the strain on the user to formulate a *good* prompt, shifting it to the tool designer. During tool development, the *prompt template* to generate ideas needs to be defined. The template needs to be flexible to accept different questions for which the user might want to generate ideas.

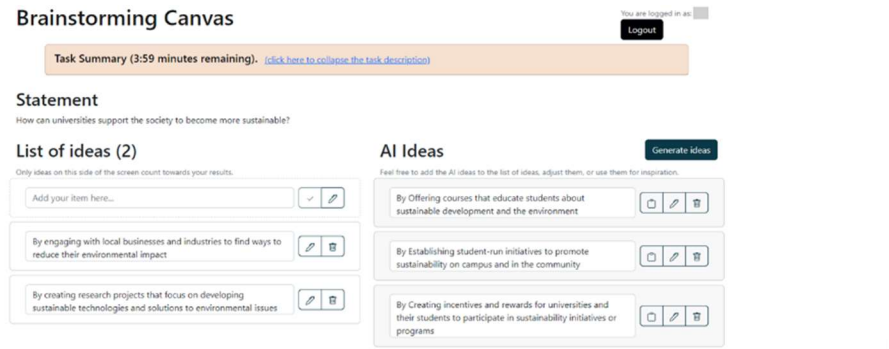


Figure 1. Screenshot of a GLM-based app prototype adapted from Memmert and Tavanapour (2023)

The sensitivity of the GLM’s output to the input (Zhao et al., 2021) raises questions about how to effectively design a *good* prompt template for ideation purposes. While prompt engineering techniques have been explored for typical natural language processing tasks (Brown et al., 2020), there is a lack of systematic investigation on the effects of these techniques in creative tasks such as ideation. Therefore, in this study, we aim to investigate the effects of six different prompt engineering techniques on the quality of ideas generated by GPT-3 on an exemplary idea generation question. To this aim, we pose the research question:

How does prompt engineering, using a GLM like GPT-3, affect the quality of ideas generated for an ideation task?

To investigate this question, we developed a baseline prompt template and all variations of the prompt template considering the six prompt engineering techniques. We filled the prompt template with an exemplary brainstorming scenario and had GPT-3 generate ideas for each prompt. The quality of the generated ideas is assessed by three rates according to novelty and value, which are typical quality criteria for idea evaluation (Siangliulue et al., 2015). We analyzed the effects of the prompt engineering techniques and all their combinations on the idea quality. Our findings suggest that more is not necessarily merrier when it comes to prompt design for creative tasks. Combining certain prompt design techniques can negatively affect the quality of ideas generated by GPT-3.

Our study contributes to the emerging field of prompt design and aims to shed light on the potential of GLMs to support ideation. Moreover, it shows a potential path for abstracting away the complexities of directly interacting with GLMs while still benefiting from its capabilities, offering a wrapper UI, helping the user to ask questions to the GLM (Dang et al., 2022). Our paper offers the foundation for studies examining humans and AI brainstorming together, as it provides insight into a key design aspect: the underlying prompt template. This work informed a follow-up study in which the actual effect of providing AI suggestions to humans during a brainstorming session was assessed according to novelty and value of ideas (under review).

12.2 Background

12.2.1 Idea generation support

The goal of idea generation is to produce as many out-of-the-box ideas as possible to solve burning problems (Schallmo & Lang, 2020). To facilitate the creativity process, computer-based tools called Creativity Support Systems (CSS) have been developed. CSS has a long research history and focuses on different aspects of the creativity process (Przybilla et al., 2019). CSS can provide participants with cognitive or social stimulation (Pilicki et al., 2022). Different tools have been developed and evaluated, and they show varying degrees of effectiveness (Gabriel et al., 2016).

One of the main areas of focus in CSS is idea generation tools, which aim to support the creative process by providing various techniques and methods for generating ideas. These tools can have a positive impact on the creative process, including improving the quality and quantity of generated ideas (Maaravi et al., 2021). However, these tools also have some drawbacks. One of the main challenges is that they often rely on a fixed set of methods and techniques, which may limit their effectiveness. Moreover, some of the tools may require a certain level of expertise or training to use them effectively (Frich et al., 2019). Recently, GLMs have emerged as a new technology that may offer a more effective way to support the creativity process.

12.2.2 Generative language models (GLM)

GLMs are a type of machine-learning-based AI systems. GLMs are trained on large corpora and are able to predict the next word given a certain input. GPT-3 is a powerful example of such GLMs (Brown et al., 2020). Due to their generative nature, GLMs have been proposed to be used for creative applications (Gero et al., 2022). However, there are several challenges with such systems. To interact with GLMs, an input text, i.e., a prompt, needs to be formulated. This can be difficult for novices, who might be surprised by the outputs of such systems at first use (Jiang et al., 2022). Dang et al. (2022) suggest supporting users in asking questions to the model. While recent advances allow humans to freely interact with GLMs via chat (e.g., ChatGPT), and one might educate users in effectively working with GLMs, a direct interaction might not always be feasible. In such cases, the GLM capability can be embedded into a system, hidden from the users. For this, a prompt template would need to be pre-defined at development time and populated with the user input at runtime. In this study, we investigate how such a prompt template should be formulated for an idea-generation application by exploring different techniques for phrasing prompts.

With regard to using GLMs for idea generation, Di Fede et al. (2022) suggested using GPT-3 for brainstorming but have not reported data yet. Stevenson et al. (2022) used GPT-3 (earlier version) for the “alternative use test” creativity test, generating ideas for alternative uses for an object. In this study, we focus on using GLMs for a different divergent thinking task, i.e., to generate ideas for solving societal problems.

12.3 Method

12.3.1 Prompt template development

Requirements for prompt templates. We seek to develop a prompt template that can be embedded into an idea-generation app (see Figure 1). The prompt template should be *flexible* to be used for different questions, i.e., it should accept a parameter for the specific question at hand (see Figure 2). The prompt, once entered into the GLM, should produce a *list of N ideas* (instead of a free text), which can be shown as items to the user in the app. Lastly, the prompt should produce *good ideas*. The goodness of ideas can be operationalized or measured according to different criteria. For this study, we use *novelty* and *value* of ideas with a definition adapted from Siangliulue et al. (2015), as these reflect common evaluation criteria used in brainstorming research (Althuisen & Reichel, 2016; Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023). *Novelty* will be rated by considering how *novel*, *original*, or *surprising* the idea is. *Value* will be judged by considering how *useful* and *practical* the idea sounds. We have included six prompt engineering techniques to investigate their effect on idea quality.

Baseline. First, we developed a baseline prompt. We tested if we could provide the question to the GLM directly without further input. Note that squared parentheses indicate a placeholder filled at runtime.

[question]

Prompting the GLM directly with the brainstorming question, however, resulted in inconsistent responses for certain questions (e.g., “How can generative language models be used?”). The results sometimes were lists of ideas (as intended), but sometimes free text, unfit for our app. Thus, we added a description of the task we expect the GLM to perform (Reynolds & McDonnell, 2021), i.e., ‘Provide 3 ideas’. To receive a bullet-point list easily processible by our app, the technique of “itemization” might be used (Mishra, Khashabi, Baral, Choi, & Hajishirzi, 2022) by indicating a starting enumeration. In this way, the GLM “recognizes” the desired output to be a list. This template meets the first two formal requirements, as it is *flexible* to be used for different ideation questions and provides a *list of ideas*.

Provide 3 ideas for the question below.

[question]

1.

In the following, we develop prompt templates based on prompt engineering techniques, applying them to the idea generation context to assess differences in the goodness of the results. All selected prompt engineering techniques might be used in creative contexts and require adding text to the baseline prompt template described above. The techniques fundamentally differ in their dependency on the specific brainstorming question. Three of the techniques are independent of the ideation question, while the others are question-specific. As a result, to implement the former techniques, changes only occur in the backend,

whereas for the latter, the users would need to enter additional information (i.e., the GUI would need to be adapted). Table 1 offers an overview of the techniques with their respective characteristics; Figure 2 shows how the type of technique affects information flow.

Table 1. Prompt engineering techniques and characteristics

Ideation question dependency	Additional User Input	Prompt Engineering Techniques
Independent	No	Evaluation criteria specification Instructions/Schema Demonstrations – question independent
Dependent	Yes	Providing context Expert perspective Demonstrations - question dependent

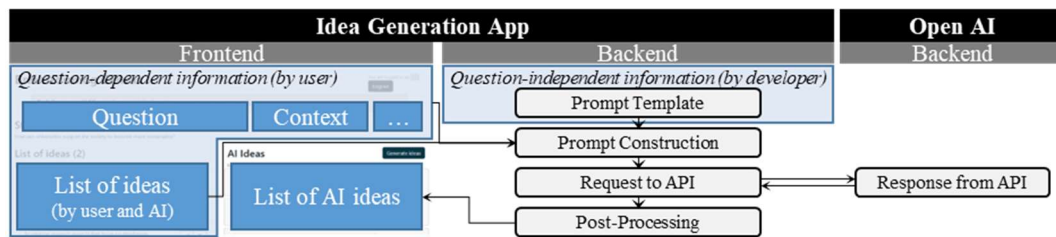


Figure 2. Illustrative depiction of the dynamic population of the prompt template with question-dependent information by the user to dynamically generate suggestions (via an API call)

Evaluation criteria specification (ECS). Specifying expectations is an important part of prompting GLMs. For brainstorming, we seek *good* ideas according to the two dimensions novelty and value with adapted definitions from Siangliulue et al. (2015), as these include common evaluation criteria for brainstorming ideas. Explicitly including expected idea characteristics in the prompt was done before (e.g., Stevenson et al., 2022) and can improve result quality (Summers-Stay et al., 2023). Thus, we included these criteria in the prompt (difference to baseline in blue):

Provide 3 *novel and valuable* ideas for the question below. *Novelty will be rated by considering how novel, original or surprising the idea is. Value will be judged by considering how useful and practical the idea sounds.*

[question]

1.

Instruction/schema (INS). Another suggested approach is to use schemas to identify the different prompt parts, e.g., by using headers (Liu et al., 2023; Mishra, Khashabi, Baral, & Hajishirzi, 2022) to label the task, question, and expected answer. This could be particularly relevant once more prompt techniques and, thereby, more information is added to the prompt.

Task: Provide 3 ideas for the question below.

Question: [question]

Answer:

1.

Demonstrations – question independent (DQI). Providing examples within the prompt can help to calibrate the GLM to better “understand” the expected results. This technique is commonly referred to as demonstration or few-shot learning (as opposed to zero-shot learning), builds on the idea of analogical reasoning (Q. Zhu & Luo, 2022) and was shown to be effective in some cases (Brown et al., 2020; Liu et al., 2023). For the area of creativity and design, Q. Zhu and Luo (2022) used design competition questions and winning ideas from previous years as examples to then generate new ideas. Inspired by this, we used pairs of brainstorming questions and good (scoring high for novelty and value) ideas published in prior studies (Nelissen, 2022; Siangliulue et al., 2015). We randomized the order of the selected ideas (demonstrations) to prevent sequential order effects (Zhao et al., 2021).

Provide 3 ideas for the question below.

What are ideas for new sports equipment products for the student market?

1. Headphones that choose which song to play aligned with your bpm (so intense music for running/ cardio and chill music for yoga/cooling down)
2. AI analytics for evaluating form from video recording
3. A device analyzing blood to indicate which supplements and vitamins are needed

Provide 3 ideas for the question below.

< question 2 - shortened here >

1. < question 2, idea 1 - shortened here >
2. < question 2, idea 2 - shortened here >
3. < question 2, idea 3 - shortened here >

Provide 3 ideas for the question below.

< question 3 - shortened here >

1. < question 3, idea 1 - shortened here >
2. < question 3, idea 2 - shortened here >
3. < question 3, idea 3 - shortened here >

Provide 3 ideas for the question below.

[question]

1.

Besides these prompt engineering techniques independent from the question, there are techniques depending on the specific question, such as adding context, expert perspective, and question-dependent demonstrations. While the former techniques aimed at improving the output quality independently from the input, these techniques adjust the input (i.e., would require additional input from the user and an adjustment of the GUI, e.g., by adding additional input fields).

Providing context (CON). GLMs generalize their outputs across the different contexts present in the training data. Thus, providing the *intended* context might improve the results to be tailored to the desired outputs (Liu et al., 2023). Besides using examples (few-shot

learning), this can be achieved by adding additional information in a zero-shot learning (i.e., no examples) setting (Brown et al., 2020). In our tool, this could be reflected through an additional ‘context’ text field for the user to add context-related information.

Provide 3 ideas for the question below.

[context]
[question]
1.

Expert perspective (EXP). Similarly to adding additional context, asking the GLM to take a certain perspective or angle might improve the output quality. Reynolds and McDonnell (2021) suggest prompting the GLM to take the perspective of a public figure or a specific role (e.g., teacher) to produce particular results; Haase and Hanel (2023) suggest including a profession. Similarly, we ask the GLM to take the role of an expert for a specific field, which is to be specified by the user.

Assuming you are an expert for [expert field], please provide 3 ideas for the question below.
[question]
1.

Demonstrations – question dependent (DQD). Previously, we explained that examples or “demonstrations” could be used to improve the GLMs' performance (few-shot learning). While the examples in the earlier technique were independent of the specific brainstorming question, once the user entered their first ideas, these ideas can be added dynamically to the prompt. For our example, we assumed the user had already added at least three ideas, and three of these ideas are included in the prompt. We randomized the idea order to prevent ordering effects (Liu et al., 2023; Zhao et al., 2021). In actual use, the best ideas might be selected algorithmically (Summers-Stay et al., 2023).

Provide 6 ideas for the question below.

[question]
1. [idea 1]
2. [idea 2]
3. [idea 3]
4.

The six selected techniques are independent of one another, allowing for any combination (i.e., all combinations are feasible). The goal of our analysis is to increase the understanding of which (combinations of) techniques lead to novel and valuable ideas. We developed a Python script to create all 64 prompt templates consistently, representing all combinations of the six prompt techniques described above (2^6 combinations from including none to including all techniques; no duplications). We recursively applied the techniques, i.e., when ‘Context’ and ‘Demonstrations – Question Independent’ were included, then context was also added to the demonstrations. Our templates are *prefix prompts* as the entire prompt text proceeds the expected output (Liu et al., 2023).

12.3.2 Idea generation

To generate ideas for our experiment, we developed an exemplary brainstorming scenario. We used the societal problem of food waste (see Table 2). Open-ended problems for which ideas can be generated with relatively common knowledge are frequently used in brainstorming studies (Y. Zhu et al., 2020, 2021); the specific societal problem of avoiding food waste was adapted from Y. Zhu et al. (2021). While the question was copied, we added the question-specific details for context and expert perspective. For the example ideas, we used ideas that scored high in novelty and value from the same prior study. We filled all 64 templates with the contents for this specific scenario. Two additional scenarios we had included originally to increase the robustness of our results we had to abandon (as will be described in the idea evaluation section).

Table 2. Exemplary scenario

Component	Text
Context	Food waste is a major issue that affects both the environment and the economy. Globally, it is estimated that about one-third of all food produced is lost or wasted each year. Food waste is a huge source of greenhouse gas emissions and wasted natural resources, and therefore – reducing food waste could help to reduce global greenhouse gas emissions, establish food security, and encourage healthy food systems.
Question	How can we reduce food waste?
Expert Perspective	sustainability and environment
3 Examples (Demonstrations – question dependent)	• Reward grocery stores that donate their near-expired food to a food-bank where people can cook this donated food • Decompose leftovers and expired food, and supply them as fertilizers to local farmers and gardeners • Create an app that allows people to offer their leftovers or extra food to those who need it

Populating the baseline prompt template resulted in the prompt below (scenario-specific content in orange):

Provide 3 ideas for the question below.
 How can we reduce food waste?
 1.

We used a custom Python script to populate all templates and make the requests to the OpenAI-API. We used the most powerful model, ‘davinci-003’, at standard settings but adjusted the temperature to 0.9, as per the documentation for creative applications. We did not make any changes to the model to improve the results or adjust them to the exemplary scenario. We post-processed the results, removing any enumeration and cutting off the ideas after the first dot, as for some ideas, added explanations made the ideas very long. This resulted in 192 ideas (3 ideas per 64 prompts). No ideas were flagged by OpenAI’s content moderation.

12.3.3 Idea evaluation

In an approach similar to Siangliulue et al. (2015), we generated random sets of about 25 ideas in random order for each scenario and had crowd workers assess them according to

novelty and value. We provided the definition for both criteria in accordance with Siangliulue et al. (2015), with *novelty* as ‘consider how novel, original or surprising the idea is’ and *value* as ‘consider how useful and practical the idea is’. We recruited participants on the Prolific platform, restricting to English-speaking individuals with a high approval rating. We aimed at 3 ratings per idea (Siangliulue et al., 2015). We included 3 randomly selected ideas all raters had to rate, enabling us to calculate inter-rater agreement. Though most participants passed the attention check items, we decided not to use the crowd-worker evaluation data due to poor data quality, i.e., poor inter-rater agreement.

Consequently, we had three blind-to-condition raters (one design thinking expert and the first two authors) independently rate all 192 ideas for one scenario on both dimensions. Having a set of judges rate all ideas is common in brainstorming research (e.g., Althuizen & Reichel, 2016; Y. Zhu et al., 2021), so is rating individual instead of sets of ideas (e.g., Haase & Hanel, 2023; Stevenson et al., 2022). We decided to include only one scenario, allowing for testing all techniques and their combinations while reducing potential fatigue from making many evaluations. Inter-rater agreement was 0.76 for novelty and 0.57 for value, or ‘excellent’ and ‘fair’ respectively (Cicchetti, 1994). Generated ideas included: “Develop software that would help shoppers calculate exact amounts of food they need depending on meal times and leftovers” (high novelty score), “Create an online platform for restaurants to share their surplus food with people who need it” (high value score), and “Implement “best by” labeling systems” (low novelty and low value score).

12.3.4 Data analysis

We used custom Python scripts for pre-processing, i.e., sorting the ratings for techniques and combinations. The clean data was analyzed with JASP statistics software (JASP, 2023). After performing assumption checks and confirming the normality of data distribution (Levene’s $p=.747$), one-way ANOVA was performed to assess if there is an effect of prompt engineering on idea quality (value and novelty). ANOVA was performed on all possible combinations of techniques, leading to 64 pairwise comparisons. To determine the direction of the effect, significant results were further inspected with Tukey’s post-hoc test, which also controls for the smaller number of observations in subgroups when performing pairwise comparisons. Since the variance of the individual ratings was higher than when averaged across the raters, we performed the same analysis on a dataset with averaged scores. Lastly, we checked whether prompt length affects the quality of the output with correlation analyses. We chose Spearman’s rho coefficient as the prompt length variable did not satisfy the normality assumptions.

12.4 Results

This study investigates the impact of using single and combined prompting techniques on idea novelty and value for a specific brainstorming question. Regarding single techniques,

the results showed that the use of Evaluation Criteria Specification had a positive and significant impact on the novelty of ideas ($F=18.250$, $p<.001^{***}$, Figure 3-a), while Context had a negative impact compared to no technique ($F=5.464$, $p=.02^*$, Figure 3-b). The use of the other single techniques did not yield any statistically significant results.

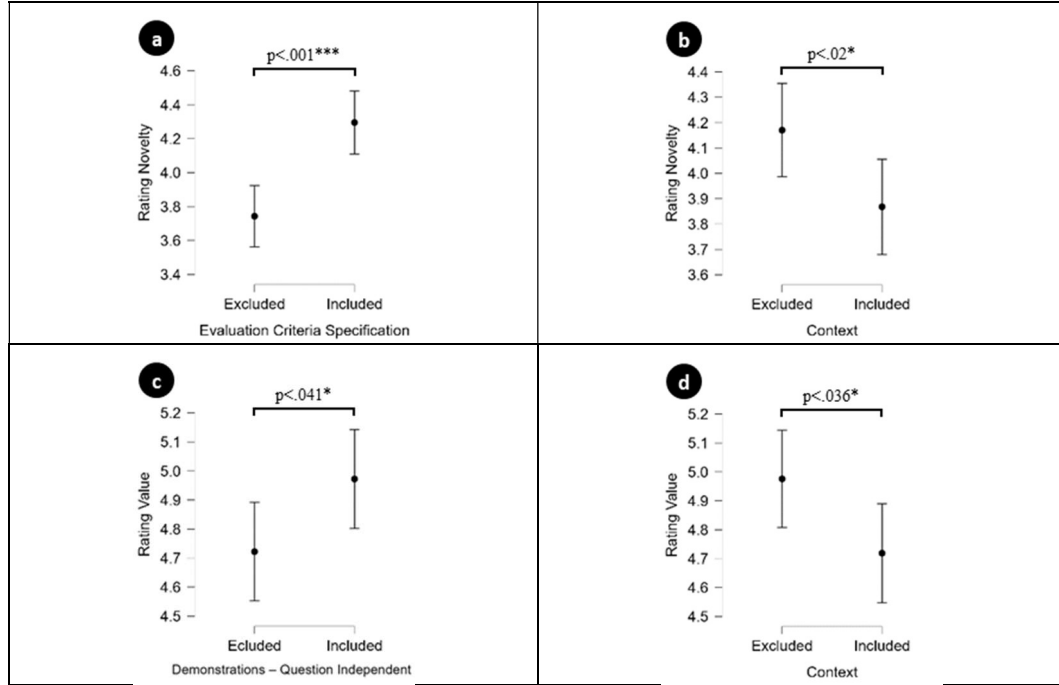


Figure 3. Techniques performing better than baseline (left) and worse than baseline (right)

Analyzing combinations of two techniques, the combination of INS and DQD showed a statistically significant effect ($F=4.736$, $p=.03^{**}$); however, the post-hoc analysis did not reveal any significant differences between groups.

Combinations of three techniques: The combination of ESC, INS, and DQD showed a statistically significant effect ($F=5.216$, $p=.023^*$), with post-hoc analysis indicating that ESC was better than INS ($t=3.815$, $p=.004^{***}$) and DQD ($t=4.460$, $p<.001^{***}$). ESC alone performed better than the combination of ESC and DQD ($t=3.332$, $p=.021^*$) and INS and DQD ($t=4.030$, $p=.002^{**}$). The combination of CON, DQI, and EXP showed a statistically significant effect ($F=8.894$, $p=.003$), with post-hoc analysis indicating that combining DQI with EXP leads to more novel ideas than CON alone ($t=-3.439$, $p=.015^*$).

The combination of ESC, CON, DQD, and EXP showed a statistically significant effect ($F=4.736$, $p=.03^{**}$), with post-hoc analysis indicating that CON alone performed worse than the combination of ESC and CON ($t=-3.572$, $p=.033^*$). The combination of ESC and CON was also found to be better than CON and DQD ($t=3.648$, $p=.026^*$).

The combination of five techniques: ESC, INS, CON, DQD, and EXP showed a significant effect ($F=8.577$, $p=.003^{**}$), with post-hoc analysis indicating that ESC was better than the

combinations of CON and DQD ($t=4.084$, $p=.018^*$); INS and EXP ($t=3.869$, $p=.04^*$), and ESC was better than the combination of ESC, INS, CON, and EXP ($t=4.084$, $p=.018^*$).

When it comes to value, the results indicate that few single techniques had a significant effect on value, with DQI performing better than when no technique was used ($F=4.194$, $p=.041^*$, Figure 3-c) and CON performing worse than baseline ($F=4.430$, $p=.036^*$, Figure 3-d). Combinations of two techniques showed that the combination of DQD and DQI had a significant effect on value ($F=9.437$, $p=.002^{**}$), with DQI performing better than no technique ($t=-3.620$, $p=.002^{**}$) and better than the combination of DQD*DQI ($t=2.735$, $p=.033^*$). The combination of DQD and EXP also showed a significant effect on value ($F=5.440$, $p=.02^*$), but post-hoc analysis did not reveal any significant differences. Combinations of three, four, and five techniques did not show significant effects on value. Despite significant effects after ANOVA, post-hoc tests did not yield significant differences ($p>.05$).

We also tested the effects of the techniques on a dataset averaged across raters to account for variance in ratings. This resulted in 3 times fewer observations, leading to a higher threshold to reach significance in pairwise comparisons. Thus, only ESC showed a positive effect on novelty ratings ($F=7.588$, $p=.007^{**}$).

We investigated the effect of prompt length on output ratings. Correlation analyses show no difference in ratings based on prompt length ($\rho=.053$, $p=.465$).

12.5 Discussion

12.5.1 Answer to the research question

This study investigated the impact of using various prompting techniques on the novelty and value of ideas generated by GPT-3. The techniques tested included Evaluation Criteria Specification (ESC), Context (CON), Instruction/Schema (INS), Demonstration: Question Dependent (DQD), Demonstration: Question Independent (DQI), and Expert Perspective (EXP). The results showed that ESC had a positive impact on the novelty of ideas, while Context had a negative impact. Our study seems to align with previous findings of Summers-Stay et al. (2023), which suggest that idea quality may be improved by adjusting the prompt. They speculate that such systems may have picked up on creative ideas for other brainstorming questions in the training data and point to the capabilities of such models for analogical reasoning. Our prompt with ESC asking for “novel ideas” might then trigger the GLM to produce such ideas. Given the complexity and opacity of such models, the mechanisms for producing such results, however, remain speculative and require further testing. Combinations of two or more techniques showed significant effects on novelty, with some techniques working better together than others. However, a few single techniques had a significant impact on the value of ideas, and combinations of techniques did not show any significant effects on value. The study's findings suggest that using specific combinations

of prompting techniques in idea generation with generative AI can enhance the novelty of ideas generated, but further research is needed to determine their impact on idea value. This further aligns with earlier findings of Summers-Stay et al. (2023), who also only found a small difference through their more advanced approach aiming for increasing utility (i.e., a dimension of *value*). As stated before, given the complexity of such models and the scarcity of empirical research (Liu et al., 2023), one can only speculate as to the reason.

Additionally, the results showed that Evaluation Criteria Specification (ESC) performed better when used alone rather than in combination with other techniques. This finding suggests that adding many techniques to the prompt may not necessarily enhance idea quality (over-engineering). A potential explanation might be “spurious correlations” (Brown et al., 2020, p. 7), a problem that can occur when adding information beyond the task (e.g., examples). Instead, it may be more effective to use a targeted approach and select a few specific techniques based on the desired outcome. This highlights the importance of prompt engineering for creative tasks, which may require a different approach than other use cases where combinations of prompts can improve results (Mishra, Khashabi, Baral, & Hajishirzi, 2022; Wu et al., 2022). Therefore, it is important to carefully consider the use of multiple techniques and assess their impact on idea quality based on the specific context and goals of the task.

Our findings of only a few statistically significant differences regarding the use of prompt engineering techniques are somewhat surprising, contrary to anecdotal evidence from social media stressing the importance of prompt engineering. One reason why we find only a few differences between the prompts could be that we produce relatively short output in a restricted format (i.e., 3 bullet points as opposed to large blocks of free text). We observed many variations of the seemingly same idea. Potentially, there are some “obvious” answers that are provided first, essentially “overshadowing” prompt differences. However, this is speculation, and further research is required, e.g., using *calibration*, as discussed by Liu et al. (2023). Other potential reasons for us observing only a few significant differences are discussed in the limitations.

12.5.2 Contribution to theory

From a theoretical perspective, the study adds to the growing body of literature on using GLMs productively in creative applications (Gero et al., 2022; Shakeri et al., 2021; Yuan et al., 2022), particularly for the case of generating ideas in brainstorming-like settings (Di Fede et al., 2022; Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023). Such idea-generation approaches typically rely on a specific, pre-defined prompt template, which may (or may not) include user-generated ideas. With this study, we enhance the understanding of the role of prompt engineering (Mishra, Khashabi, Baral, & Hajishirzi, 2022; Wu et al., 2022). Such an understanding is important, as there is an increasing number of products that embed GLMs. For such systems, a crucial aspect of system design is

to formulate an appropriate prompt template. In doing so, system designers essentially help the user to ask the GLM “the right questions” (Dang et al., 2022).

By examining the effect of different single and combined techniques on both idea novelty and value, we enhance the understanding of how to design systems for effective ideation with GLMs. The finding that ESC performed better when used alone rather than in combination with other techniques challenges the assumption that combining many techniques enhances the output quality. It rather highlights the need to consider using multiple techniques carefully in creative tasks and the importance of prompt engineering.

On a more abstract level, we contribute to the field of CSS by investigating the next-generation GLM-driven systems for supporting human creativity. Such a GLM-based brainstorming system was also suggested by Di Fede et al. (2022); with our analysis, we can inform the underlying technical design.

12.5.3 Implications for practice

From a practical perspective, the study offers valuable insights for organizations seeking to enhance their innovation potential. By identifying which techniques are most effective in enhancing the idea novelty and value, organizations can design more effective ideation sessions and improve their creative output. The finding that ESC has a positive impact on novelty highlights the importance of clearly defining brainstorming evaluation criteria to shift the focus on specific goals to generate more creative ideas. Additionally, the finding that combining DQI with EXP leads to more novel ideas than CON alone has practical implications for the use of expert perspectives in idea generation. By combining the expertise of internal and external experts with demonstrational prompts, organizations can generate more innovative ideas. However, results are to be interpreted carefully, as our study so far only indicates potential trends, but no general conclusions can be drawn.

There are large differences in the length of the prompts (14 to 614 words), particularly due to the *question-independent demonstrations* or *context* techniques. When costs for GLM usage are based on prompt length, this can result in cost differences for the prompt templates, which should be considered by tool designers, given the small differences in output quality.

12.5.4 Limitations & outlook

With only a few statistically significant differences between prompt engineering techniques, our results run counter to our (and many anecdotal) expectations. However, the results are to be interpreted carefully due to several limitations of our study that we discuss below.

While we based the candidate prompt development on existing literature, there are many ways of operationalizing the different prompt engineering techniques, and different implementations may result in different outputs. We, however, only tested one operationalization

per technique, prioritizing testing more techniques over more operationalizations per technique. Additionally, the field of prompt engineering is rapidly evolving, and new techniques are constantly suggested. Thus, future research may expand on our results for both adjusting the operationalization of techniques or including additional techniques.

We have selected GPT-3 because it is a powerful, widely adopted model (Brown et al., 2020). However, there are other GLMs, and GLMs are constantly evolving, which, in some cases, affects the quality of GLM-generated ideas (Haase & Hanel, 2023). Additionally, there is a dependency between the effectiveness of prompts and model sizes (Liu et al., 2023), which could affect prompt techniques and transferability of results; some even suggest prompt engineering might become obsolete when models improve (Oppenlaender et al.). Thus, future research should explore how different model (versions) together with prompt techniques affect the idea quality.

Having 64 templates, each technique was present in 32 prompts. With 3 ideas per prompt, we had 96 ideas with and without each prompt technique, allowing for meaningful comparisons. However, including combinations of techniques reduced the number of observations per group. Tukey's post-hoc test takes into account the variability of the data and sets a threshold for statistical significance that is adjusted for the number of pairwise comparisons being made, e.g., if there are 16 pairwise comparisons, the adjusted p-value is $0.05/16=0.003$, thus making it harder to spot significant differences. To counteract this, future studies could rely on these results to investigate fewer techniques and additionally collect a higher number of observations.

Originally, we had planned to assess the robustness of our results across three societal problem questions; however, due to the low level of agreement among the crowd workers' responses, we had to use a panel of three raters to rate all ideas. To avoid fatigue, ideas for only one scenario were assessed (192 ideas on two dimensions). We thereby prioritized testing technique combinations and high data quality over additional scenarios. Thus, we cannot report insights on the robustness across problem questions. While the nature of the assessed information system – a general-purpose GLM without adaptations to our scenario – conceptually does not give any particular reason to assume that it would function completely differently across scenarios that are relatively common knowledge, future research should investigate these results to increase robustness.

For assessing idea quality, we selected two evaluation criteria (novelty, value) that reflect the criteria commonly used in literature. However, there are other criteria, such as practicability or plausibility of ideas (incl. costs), as well as criteria for sets of ideas such as *diversity* (Siangliulue et al., 2015), or even more question-specific criteria, such as social utility or moral value. Future research should further explore how GLMs could support idea generation along such criteria.

Lastly, inter-rater agreement for the three-rater panel was higher than for crowd-workers. However, an even better agreement, particularly for ‘value’, might have surfaced more statistically significant findings.

Given these limitations, we call for further investigation of different ways of operationalizing the prompt engineering techniques and an application to other problems to increase robustness. Our approach is transferable to other brainstorming questions and can serve as a foundation. Additionally, we suggest increasing inter-rater agreement, e.g., by increasing the number of raters or by using an evaluation scheme. More broadly, the ideation performance of humans working with such systems should be investigated.

While not at the core of our study, future research needs to investigate the implications of GLMs' inability to *understand* language and develop mitigation strategies for the issues of biases, falsehoods, and lack of moral judgment – discussed more broadly for GLMs (Bender et al., 2021; Floridi & Chiriatti, 2020; Lin et al., 2022; Susarla et al., 2023) – for creativity- and work-related use, particularly as brainstorming sessions are framing problem and solution corridors. Depending on the specific problem at hand, biased GLMs' suggestions could have severe negative consequences. We thus urge future research to explore sociotechnical perspectives of integrating generative AI in work settings by leveraging human expertise appropriately and preventing humans from accepting unfit suggestions, potentially inspired by existing research on the engagement of humans with AI systems' outputs from the area of AI-assisted decision-making (e.g., Buçinca et al., 2021).

12.6 Conclusion

GLM capabilities may be embedded into products, abstracting away the difficulties of interacting with the GLM directly while enabling unskilled users to still benefit from it. While this frees the user from developing a *good* prompt, it makes the design of a good *prompt template* during system development necessary. In our study, we develop a set of 64 prompt templates according to six prompt engineering techniques for an ideation tool, generate ideas via the GLM ‘GPT-3’ for an exemplary societal problem, and have these ideas evaluated. We find that prompt engineering techniques only in a few cases significantly affect idea quality positively. As this is counter to popular belief and our own experience, we carefully discuss potential explanations and limitations of our study. Particularly, we encourage further research on the effects of using prompt design for creative, divergent thinking tasks.

Acknowledgment

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

12.7 References

Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of*

- Management Information Systems*, 33(1), 11–44.
<https://doi.org/10.1080/07421222.2016.1172439>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM.
<https://doi.org/10.1145/3442188.3445922>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21.
<https://doi.org/10.1145/3449287>
- Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, 6(4), 284–290. <https://doi.org/10.1037/1040-3590.6.4.284>
- Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models.
<http://arxiv.org/pdf/2209.01390v1>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Frich, J., MacDonald Vermeulen, L., Remy, C., Biskjaer, M. M., & Dalsgaard, P. (2019). Mapping the Landscape of Creativity Support Tools in HCI. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). ACM.
<https://doi.org/10.1145/3290605.3300619>
- Gabriel, A., Monticolo, D., Camargo, M., & Bourgault, M. (2016). Creativity support systems: A systematic mapping study. *Thinking Skills and Creativity*, 21, 109–122.
<https://doi.org/10.1016/j.tsc.2016.05.009>
- Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing using Language Models. In F. ` Mueller, S. Greuter, R. A. Khot, P. Sweetser, & M. Obrist (Eds.), *Designing Interactive Systems Conference* (pp. 1002–1019). ACM.
<https://doi.org/10.1145/3532106.3533533>

- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- JASP. (2023). (*Version 0.17*)[Computer software].
- Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J [Carrie J.] (2022). PromptMaker: Prompt-based Prototyping with Large Language Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems* (pp. 1–8). ACM. <https://doi.org/10.1145/3491101.3503564>
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9), 1–35. <https://doi.org/10.1145/3560815>
- Maaravi, Y., Heller, B., Shoham, Y., Mohar, S., & Deutsch, B. (2021). Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science*, 15(6), 1431–1464. <https://doi.org/10.1007/s11846-020-00400-5>
- Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/107289>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
- Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 589–612).
- Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Nelissen, E. R. (2022). *Improving individual idea generation with the selective brainwriting technique*. <https://purl.utwente.nl/essays/91388>
- Oppenlaender, J., Linder, R., & Silvennoinen, J. Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering. Advance online publication. <https://doi.org/10.48550/arXiv.2303.13534>
- Pilcicki, R., Siemon, D., & Lattemann, C. (2022). How Feature- and Communication Constraints in CSS Affect Creative Collaboration in Virtual Teams - An Activity Theory Perspective. In *55th Hawaii International Conference on System Sciences*.

- Przybilla, L., Baar, L., Wiesche, M., & Krcmar, H. (2019). Machines as Teammates in Creative Teams. In D. Joseph, C. van Slyke, J. P. Allen, J. Quesenberry, & M. Wiesche (Eds.), *Proceedings of the 2019 on Computers and People Research Conference* (pp. 94–102). ACM. <https://doi.org/10.1145/3322385.3322402>
- Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). ACM. <https://doi.org/10.1145/3411763.3451760>
- Schallmo, D., & Lang, K. (2020). *Design Thinking erfolgreich anwenden: So entwickeln Sie in 7 Phasen kundenorientierte Produkte und Dienstleistungen* (2., aktualisierte Auflage). Springer Gabler. <https://doi.org/10.1007/978-3-658-28325-4>
- Shakeri, H., Neustaedter, C., & DiPaola, S. (2021). SAGA: Collaborative Storytelling with GPT-3. In J. Birnholtz, L. Cioffi, S. Ding, S. Fussell, A. Monroy-Hernández, S. Munson, I. Shklovski, & M. Naaman (Eds.), *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 163–166). ACM. <https://doi.org/10.1145/3462204.3481771>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). *Putting GPT-3's Creativity to the (Alternative Uses) Test*.
- Summers-Stay, D., Voss, C. R., & Lukin, S. M. (2023). Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.
- Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, 34(2), 399–408. <https://doi.org/10.1287/isre.2023.ed.v34.n2>
- Wu, T., Terry, M., & Cai, C. J [Carrie Jun] (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM. <https://doi.org/10.1145/3491102.3517582>
- Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (pp. 841–852). ACM. <https://doi.org/10.1145/3490099.3511105>
- Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*. <https://proceedings.mlr.press/v139/zhao21c/zhao21c.pdf>

- Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, 2, 1825–1834. <https://doi.org/10.1017/pds.2022.185>
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2020). Creativity: Intrapersonal and Interpersonal Selection of Creative Ideas. *The Journal of Creative Behavior*, 54(3), 626–635. <https://doi.org/10.1002/jocb.397>
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2021). The effect of rank-ordering strategy on creative idea selection performance. *Eur. J. Soc. Psych.*, 51(2), 360–376. <https://doi.org/10.1002/ejsp.2743>

13 Publication 4: Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects

Memmert, L. (2024). Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects. In *Proceedings of the 32nd European Conference on Information (ECIS)*, Paphos, Cyprus. https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1

Abstract

Generative language models (GLMs) like GPT-3.5 or ChatGPT produce remarkable results when used for creative tasks. In this short paper, we outline our planned approach to investigating how humans brainstorm together with GLMs. Particularly, we plan to explore how the presentation of GLM suggestions affects brainstorming group effects such as cognitive stimulation or social loafing known from all-human groups. Based on group brainstorming literature and previous studies, we designed an experiment to measure underlying performance effects. In the planned between-subject experiment, participants brainstorm with a GLM. We investigate how manipulating the completeness and the communicated origin of suggestions affects performance-related constructs. In this short paper, we only report preliminary results. Our study has theoretical implications explaining how well-documented brainstorming group effects from all-human groups can help to understand and affect human-AI team performance. Practical implications include insights into designing GLM-based creativity support while balancing performance dimensions.

Keywords: GPT-3.5, Generative AI, Brainstorming, Cognitive Stimulation, Social Loafing

13.1 Introduction

Generative language models (GLM) like GPT-3 or ChatGPT are useful in supporting creative tasks and are capable of producing creative ideas comparable to many humans (Haase & Hanel, 2023). Going beyond an isolated comparison of humans and AI, it was suggested to investigate a joint session in which humans and GLMs brainstorm together (Di Fede et al., 2022). Multiple studies have investigated the effect on performance, suggesting that working with a GLM might be associated with higher brainstorming performance (e.g., Dell'Acqua et al., 2023; Memmert & Bittner, 2024).

Conceptually, by introducing a GLM (i.e., an AI), individual brainstorming of a human is turned into collaborative brainstorming of a human and AI and, thus, an instance of human-AI collaboration. An important research challenge for human-AI collaboration is to develop an understanding of and improve the performance of human-AI teams (Dellermann et al., 2019).

When humans brainstorm as part of a group, certain mechanisms that increase or decrease brainstorming performance can occur (Pinsonneault et al., 1999). We refer to those mechanisms as *brainstorming group effects*. In this study, we propose to use *brainstorming*

group effects from all-human brainstorming groups as a lens to improve the understanding of the brainstorming performance of human-AI teams. Additionally, we propose to use *cognitive forcing*, an approach that was applied in AI-assisted decision-making (Bućinca et al., 2021), as a potential means for affecting and improving human-AI team performance. Research on all-human brainstorming groups has uncovered several group effects affecting individuals' brainstorming performance (Pinsonneault et al., 1999). As humans might socially respond to technical systems (Nass & Moon, 2000), it was argued that such group effects of all-human settings might occur in human-AI settings as well. Earlier work with less advanced AI(-like) systems indeed found different manifestations of those known effects in human-AI teams (Siemon et al., 2015; Stieglitz et al., 2022). Recent research on human-GLM brainstorming (Memmert & Tavanapour, 2023) hinted at humans feeling inspired by the GLM to develop additional ideas (performance-enhancing *cognitive stimulation*) but potentially reducing their effort when working with the GLM (*performance-reducing social loafing*). To better understand how these group effects impact brainstorming performance, we propose an experiment based on cognitive forcing. Cognitive forcing refers to interventions seeking to have humans engage more thoughtfully with AI suggestions (Lambe et al., 2016). They have been used in AI-assisted decision-making, i.e., a different area of human-AI collaboration (Bućinca et al., 2021; Lambe et al., 2016). We propose to test transferring this approach onto our creative brainstorming task, varying how information of the GLM is displayed to understand the effect on performance-influencing aspects, seeking to answer the research question:

How will the completeness of GLM suggestions in a human-AI brainstorming setting affect the brainstorming performance through the brainstorming group effects of cognitive stimulation and social loafing?

Brainstorming Canvas

You are logged in as: 213

Logout

Task Summary (3:59 minutes remaining). [\(click here to collapse the task description\)](#)

Statement

How can universities support the society to become more sustainable?

List of ideas (2)

Only ideas on this side of the screen count towards your results.

Add your item here...

By engaging with local businesses and industries to find ways to reduce their environmental impact

AI Ideas

Generate ideas

Feel free to add the AI ideas to the list of ideas, adjust them, or use them for inspiration.

By Offering courses that educate students about sustainable development and the environment

By Establishing student-run initiatives to promote sustainability on campus and in the community

Figure 1. Brainstorming prototype with statement, user input, and AI suggestions

To investigate this question, we have developed a prototype (Figure 1), allowing users to submit ideas and receive inspiration from a GLM. We suggest a 2x2 between-subject design, exploring how presenting the GLM suggestions affects brainstorming performance-related group effects. While the data collection is still in progress and thus does not yet allow for a comprehensive analysis, we report initial results as part of this short paper, using data from an earlier study with the same prototype.

With our study, we aim to contribute to theory by shedding light on whether selected, documented group effects from all-human brainstorming (Pinsonneault et al., 1999) can be transferred to human-AI brainstorming groups (as was discussed, e.g., by Stieglitz et al., 2022), and whether cognitive forcing interventions inspired from AI-assisted decision-making (Bućinca et al., 2021) can be transferred into creative contexts. We thereby seek to improve the understanding of performance in human-AI teams and lay the foundation for improving the performance of human-AI teams in creative settings. The tested interventions might inform the sociotechnical design of (individual) creativity support systems to induce performance-enhancing group effects while minimizing performance-reducing effects.

13.2 Background

13.2.1 Brainstorming Group Effects

Brainstorming originally referred to a creative technique in which a group of humans develops ideas (Osborn, 1953). Over the decades, researchers explored various brainstorming group sizes, from solitary brainstorming without interacting with others to brainstorming in (large) groups (Pinsonneault et al., 1999). Exploring how being part of a group affects individuals' brainstorming performance, several performance-enhancing and -reducing mechanisms (we here refer to them as brainstorming group effects) have been identified (Latané et al., 1979; Pinsonneault et al., 1999), such as *cognitive stimulation* and *social loafing*. We here focus on cognitive stimulation and social loafing, as previous research found hints for their occurrence in human-AI teams (Memmert & Tavanapour, 2023; Siemon et al., 2015; Stieglitz et al., 2022). Cognitive stimulation refers to the performance-enhancing effect that “utterance of members may contain task related stimuli that elicit new ideas from other members” (Pinsonneault et al., 1999, p. 114). Thus, when team members brainstorm together, they benefit from each other's ideas, which might help them in developing additional ideas. However, for such cognitive stimulation to occur, group members need to pay *attention to each other's ideas* (Dugosh et al., 2000). Such attention to others' ideas, however, might negatively affect performance by causing the “distraction effect”, as it reduces time to think up and express own ideas (Pinsonneault et al., 1999, p. 126). Social loafing and free riding are closely related concepts referring to the performance-reducing mechanism of individuals reducing their effort when working in groups as compared to when working alone (Latané et al., 1979; Pinsonneault et al., 1999). Reasons why humans might reduce their effort when working in a group include, e.g., (1) *perceived dispensability of effort* and (2) *diffused responsibility* (Pinsonneault et al., 1999). Perceived dispensability refers to individuals feeling their contribution would not make (much) of a difference. Diffused responsibility refers to humans feeling less personal responsibility. Both effects can lead to individuals contributing less (Pinsonneault et al., 1999). Social loafing is considered negative, as it comes at other group members' expense, who have to compensate for the effort reduction.

Brainstorming group effects, i.e., mechanisms like cognitive stimulation or social loafing, positively or negatively affect brainstorming performance. While traditionally observed in purely human teams, several studies investigated whether these effects known from all-human groups also occur when working with non-human team members (Siemon et al., 2015; Siemon & Wank, 2021) or even theorized that humans might reduce their effort when working with non-human team members (Liu et al., 2023; Stieglitz et al., 2022). GLMs can now produce creative ideas comparable to humans (Dell’Acqua et al., 2023; Haase & Hanel, 2023). Such systems might, thus – comparable to humans – induce cognitive stimulation by offering inspiration. However, due to their high performance, there might be a higher chance of social loafing. In this study, we explore whether the known brainstorming group effects from all-human groups are useful for understanding brainstorming performance in human-GLM teams. To understand the differences between human and AI teammates, one of the manipulations will focus on the communicated origin of the ideas (human vs. AI), though the AI will generate all ideas.

13.2.2 Cognitive Forcing

In settings of joint work between human and AI, the presentation of AIs’ output affects how humans perceive and work with such AIs’ outputs. One of the challenges that arise in such settings is that humans might not engage with AI outputs individually but instead develop heuristics about the general competence of the AI when working with it (Buçinca et al., 2021). “Cognitive forcing functions” could be a means to prevent such behavior. Cognitive forcing functions refer to “interventions that are applied at the decision-making time to disrupt heuristic reasoning and thus cause the person to engage in analytical thinking” (Lambe et al., 2016). Inspired by such interventions used in AI-assisted decision-making (Buçinca et al., 2021), we suggest transferring them into our creative context when working with GLMs by altering the output of the GLM at presentation time. For our experiment, we suggest two GLM output types: (1) complete ideas (i.e., comparable human ideas) and (2) ideas shortened to just a few words. We plan to assess if the intervention increases the human’s attention to the GLM suggestions.

13.2.3 Hypothesis Development

When participants brainstormed with a GLM, hints for both performance-enhancing and -reducing effects surfaced (Memmert & Tavanapour, 2023). We plan to investigate the effect of two modes of presenting GLM suggestions on brainstorming performance. By manipulating suggestion presentation, we seek to increase the brainstorming performance-enhancing and decrease the performance-reducing effects. We derive the hypotheses of the conceptual model with individual hypotheses (Figure 2) below.

When the GLM provides complete ideas comparable to humans, participants do not have to engage critically with these GLM idea suggestions but can merely copy the suggestions. To counteract this behaviour, we suggest an intervention inspired by the cognitive forcing approach used in AI-assisted decision-making (Buçinca et al., 2021; Lambe et al., 2016).

We propose as a manipulation to reduce GLM suggestions from complete ideas to just hints of a few words. We hypothesize that when suggestions are only partial, humans would need to engage more carefully with the suggestions, which could increase their level of attention to suggestions (H1a+). Such heightened attention to others' ideas is critical for cognitive stimulation to occur (Dugosh et al., 2000). We therefore propose that attention to others' ideas will be positively correlated with cognitive stimulation (H1b+). Cognitive stimulation typically manifests through higher individual brainstorming performance (H1c+, H1d+). While attention to others' ideas is necessary for cognitive stimulation, spending time reading those ideas could prevent the human from expressing their own ideas ("distraction effect", Pinsonneault et al., 1999), which could reduce their individual performance (H1e-).

GLMs can provide ideas similar to humans (Haase & Hanel, 2023). In this setting, social loafing could manifest as the human "simply" accepting ideas without revision. Our proposed manipulation (reduce GLM suggests to just a few words) counteracts this behaviour as humans cannot social loaf by simply accepting the (incomplete) hints into the final list of suggestions because *complete* ideas are expected as a result. Hence, we propose social loafing to be negatively correlated with the number of human ideas, and positively with the number of AI ideas (H2e-, H2f+). Stieglitz et al. (2022) suggested that humans working with non-human agents might save cognitive resources by relying on the non-human teammate. We thus propose a negative correlation between social loafing and cognitive load (H2g-).

Humans might reduce their effort (i.e., social loaf) when they perceive their effort to be *dispensable* (H2c+; Pinsonneault et al., 1999), which is particularly likely to occur for less able individuals (Price et al., 2006). Additionally, *diffused responsibility* might cause social loafing (H2d+; Latané et al., 1979), which might occur given that the GLMs can contribute complete ideas similar to a human. We propose that by providing only partial suggestions, the effort of the human becomes more important (i.e., less dispensable, H2a-), and responsibility is shifted back to the human (i.e., less diffused, H2b-).

Group effects are known from brainstorming research in *all-human* groups. However, humans are known to socially respond to technical systems (Nass & Moon, 2000), and brainstorming performance can be affected by the brainstorming partner being a bot or a human (Hwang & Won, 2021). It is, however, unclear if those brainstorming group effects apply to human-AI teams (Hwang & Won, 2021) in a similar fashion. We thus investigate the effect of having the suggestions presented as originating from a human or a GLM (difference in *suggestion source*; undirected hypotheses).

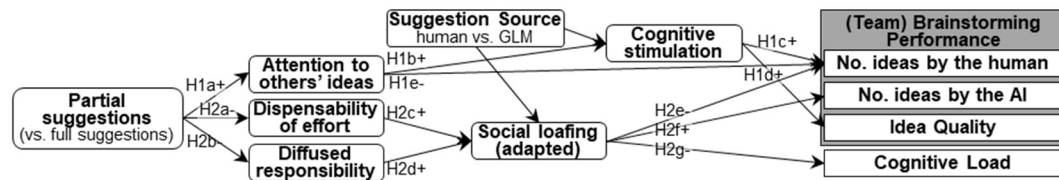


Figure 2. Conceptual model with brainstorming group effects and related concepts

13.3 Method

We developed an app (Figure 1) based on the GLM ‘GPT-3.5’, which was used for the pre-study and the planned study. The app allows users to enter, edit, and remove ideas and to request, view, edit, or copy GLM suggestions. In the backend, the prompts are constructed automatically and sent to the GLM, which provides suggestions to be shown to the human.

13.3.1 Pre-study

In a pre-study (Memmert & Bittner, 2024), 52 students brainstormed for 10 minutes with the app on a societal problem (food waste), selected their best ideas, and completed a post-questionnaire. The study compared performance between two conditions: with GLM support (treatment) vs. without GLM support (baseline). Participants were distributed evenly across conditions. We asked participants in the post-questionnaire about their experience using existing instruments from the literature, e.g., *cognitive stimulation* (Gozzo et al., 2022); for *social loafing*, we adapted the instrument for “*social loafing* with virtual assistants” (Stieglitz et al., 2022), *dispensability of effort* we operationalized via “perceived instrumentality” (Ivanov & Cyr, 2014), and *diffused responsibility* through responsibility attribution.

13.3.2 Adaptation for planned study

The planned study follows the same procedure as the pre-study, using the same app and post-questionnaire with the same instruments. As a baseline, participants generate ideas alone, i.e., without GLM suggestions. With our manipulations, we have a 2 (GLM suggestion completeness: full vs. partial) x 2 (communicated source of the ideas: GLM vs. human) design. Thus, there is an overlap between the studies, i.e., the pre-study’s treatment condition corresponds to the planned study’s condition of ‘full’ *GLM suggestion completeness* with ‘GLM’ as the *communicated source of the ideas*. This overlap enables us to report preliminary results here. The idea quality is evaluated on novelty and value on a 7-point Likert scale by three human evaluators (Siangliulue et al., 2015). To analyse our conceptual model, we require a larger sample. Following Siangliulue et al. (2015), we will recruit English-speaking crowd workers with a high approval rate to participate online in our experiment with a completion time of about 25 minutes. For analysing our conceptual model that relates brainstorming group effects and brainstorming performance, we plan to use structural equation modelling. To test our hypotheses, in one of the conditions participants

are led to believe the suggestions originate from humans, although they are GLM-generated. Thus, institutional review board approval is required, and a debriefing is included.

13.4 Results so far

As the pre-study followed the same procedure as the planned study, we report preliminary results based on the data already collected. While the previous study used a different manipulation, i.e., (non-) presence of GLM suggestion as opposed to modes of presenting GLM ideas, it offers initial insights into brainstorming performance through group effects. Given the group-related nature of these known brainstorming group effects, we can only collect this data for the participants with GLM support (i.e., $n=26$). We performed assumptions checks and report the appropriate correlation metrics Pearson's r or Spearman's ρ (rho); Table 1 presents the results that we can already report based on the collected data.

We were interested in two effects to affect performance in our setting: the performance-enhancing effect of *cognitive stimulation* and the performance-reducing effect of *social loafing*. For cognitive stimulation to occur, participants need to pay *attention to the others'* ideas (i.e., ideas of the GLM system). We assumed that higher *attention* would be associated with higher *cognitive stimulation*. However, we did not find a significant correlation ($r=0.020$, $p=0.922$). *Cognitive stimulation* is supposed to improve the performance of group members by inspiring or stimulating them to develop new or better ideas. Consequently, we expected a positive correlation between *cognitive stimulation* and the *number of ideas by the human* as well as the idea quality (novelty and value), respectively. However, we did neither observe significant correlations for *quantity* nor *quality* of ideas with *cognitive stimulation* (quantity: $r=0.031$, $p=0.880$; quality-novelty: $\rho=-0.115$, $p=0.610$; quality-value: $r=-0.051$, $p=0.822$), i.e., no performance improvement with heightened cognitive stimulation. While *attention to others' ideas* is required for *cognitive stimulation* to occur, it can reduce the human's time to think of their own ideas (i.e., distraction effect). We thus assumed a negative correlation between *attention to others' ideas* and the *number of ideas by the human*. Indeed, we find a significant negative correlation ($r=-0.724$, $p<0.001$).

We also investigated the performance-reducing effect of *social loafing*. Based on the literature on human brainstorming groups, we discussed two causes for effort reduction: (1) *dispensability of effort* and (2) *diffused responsibility*. We operationalized (1) the *dispensability of effort* through *perceived instrumentality*. However, we did not find a significant correlation between *perceived instrumentality* with *social loafing* ($\rho=-0.048$, $p=0.816$; and neither for topic knowledge nor for creative self-efficacy with *social loafing*). For (2) *diffused responsibility*, we asked participants for their level of responsibility for the final idea list. We find a significant negative correlation with *social loafing* ($r=-0.725$, $p<0.001$).

Social loafing manifests in humans producing fewer ideas. We thus hypothesized to observe higher levels of *social loafing* to occur with a lower *number of ideas of the human*.

The correlation between *social loafing* and the *number of ideas by the human* was significant and negative ($r=-0.710$, $p<.001$). With *social loafing* (for virtual assistants), we expected a higher degree of *social loafing* to occur with a higher *number of ideas by the GLM*. We did observe a positive correlation ($\rho=0.728$, $p<0.001$). We expected that making more use of the GLM would relieve the humans, i.e., expected higher *social loafing* to decrease *cognitive load*. However, the correlation was not significant ($r=-0.122$, $p=0.552$).

Table 1. Results of the (pairwise) correlation analysis
(*** indicate a significant correlation)

Hypothesis	Construct 1	Construct 2	Correlation	Coefficient	p-value
H1b	Attention to others' ideas	Cognitive stimulation	Pearson's r	0.020	0.922
H1c	Cognitive stimulation	Quantity	Pearson's r	0.031	0.880
H1d-novelty	Cognitive stimulation	Quality-novelty	Spearman's ρ	-0.115	0.610
H1d-value	Cognitive stimulation	Quality-value	Pearson's r	-0.051	0.822
H1e	Attention to others' ideas	No. ideas by the human	Pearson's r	-0.724***	<.001
H2c	Dispensability of effort	Social loafing	Spearman's ρ	-0.048	0.816
H2d	Diffused responsibility	Social loafing	Pearson's r	-0.725***	<.001
H2e	Social loafing	No. ideas by the human	Pearson's r	-0.710***	<.001
H2f	Social loafing	No. ideas by the GLM	Spearman's ρ	0.728***	<.001
H2e	Social loafing	Cognitive load	Pearson's r	-0.122	0.552

13.5 Discussion and Implications

In this short paper, we report first insights on brainstorming performance and group effects in human-AI groups based on (pairwise) correlation analysis. Somewhat surprising and counter to participants' subjective, qualitative feedback from earlier studies (Memmert & Tavanapour, 2023), we did not observe cognitive stimulation to be correlated with human performance. While participants might have felt more stimulated, this perceived stimulation did not improve their performance. This raises the question if a core ideas of group brainstorming – to inspire each other to generate additional ideas – occurs in this setting, i.e., is the performance of the human-AI team higher than the sum of its parts?

Attention to others' ideas did not positively affect cognitive stimulation. However, it was negatively correlated with human performance, indicating that humans spend their time on GLM ideas instead of contributing own ideas. This threat of a distraction effect is a problem known from human groups (Pinsonneault et al., 1999). Here, such an effect should be addressed through appropriate sociotechnical design (e.g., 'cognitive forcing' as discussed in this study). Humans require support in finding the right balance between thinking about their own ideas and reviewing and adjusting the AI system's output.

For social loafing, we did not observe significant effects for dispensability of effort (instrumentality, creative self-efficacy, topic knowledge). Potentially, the self-reported measures

used could conceal certain effects. However, lower- and higher-performing individuals potentially benefit differently from working with GLMs (Dell’Acqua et al., 2023), which could be a potential explanation but requires further investigation. We did observe a strong correlation with regard to diffused responsibility. Here, causation is particularly unclear: did humans feel less responsible due to being partnered up with an AI (as suggested by “traditional” brainstorming literature, e.g., Pinsonneault et al., 1999), or did humans report lower levels of responsibility because of their (lower) share of contributed ideas (also see *responsibility attribution* discussion in Stieglitz et al., 2022), resulting in fewer contributions by them?

Lastly, Stieglitz et al. (2022) suggested that humans working with non-human teammates might be able to reduce their mental workload. We adapted their “social loafing for virtual assistants”-instrument for brainstorming with GLMs but did not find a significant correlation. Potentially, the cognitive load was not reduced but shifted from ideation to reviewing and editing AI ideas.

In summary, while individual human performance was not improved working with GLMs might make some humans feel more productive. Indeed, such motivational aspects are known in human brainstorming groups as well (Pinsonneault et al., 1999), and so is over-claiming one owns’ share of work. In the case of human-AI teams, however, this perception is not at the expense of other human team members. Future research should investigate such motivational aspects and how they could be used to improve performance and satisfaction with the system. Additionally, we suggest future research to investigate how humans can be supported in managing their effort investment when working together with AI, particularly in finding the right balance between creating (own content) and curating (AI outputs) to enable the worker to maintain cognitive resources while enhancing work efficiency.

More broadly, the initial findings suggest that group effects known from all-human brainstorming can be used to understand performance in human-AI teams (also see Stieglitz et al., 2022). This is important, as there already is a strong scientific foundation. However, the results also show that some modifications are required in this new setting of collaboration with powerful GLMs. Particularly, as suggested by Stieglitz et al. (2022), social loafing might need to be reconsidered to be smart in certain circumstances. Before the background of brainstorming being a creative task, it should be highlighted that recent findings suggest using GLM for brainstorming might lead to “more homogenized outputs” (Dell’Acqua et al., 2023, p. 12) and the long-term effects of using such tools on individual ability are unclear, which could have undesired side-effects and requires further research.

Our study has several limitations. Using the data from a pre-study, our sample is limited and only covers one of the conditions (suggestion *completeness*: complete AND origin: AI) suggested in this short paper, preventing us from conducting a more comprehensive analysis (e.g., Structural Equation Modeling). As described, we therefore seek to recruit a larger

sample via a crowd-working platform. Additionally, while we have selected the two brainstorming group effects based on findings from prior literature, there are many other mechanisms (Pinsonneault et al., 1999) that might affect results. Lastly, making people believe they interact with a human when they interact with an AI (one of the manipulations) is necessary to understand the difference in influence on group effects, but it might affect result validity and raise ethical concerns. We, thus, include manipulation checks and seek institutional review board approval.

13.6 Conclusion and Outlook

GLMs like ChatGPT offer new capabilities, potentially allowing humans to work more effectively in creative settings, such as idea generation. However, this implies that the individual work of a human is turned into the collaborative work of human and AI. Given that humans are known to perceive systems as social actors, this might mean that known effects within all-human brainstorming groups occur in this human-AI setting as well, potentially offering the opportunity to leverage decades of research to better understand human-AI teams through a well-established perspective.

Therefore, in this short paper, we suggest to assess performance-related group effects. We outline our planned study approach and offer preliminary insights into the effects on performance. Given the promising nature of our preliminary results, we plan to conduct our experiment as described above and to perform a more comprehensive analysis of the inter-related group effects beyond pairwise correlation analysis. With our study, we seek to contribute to the understanding of how humans can work together with GLMs more effectively and, consequently, how tools employing GLMs should be designed to support creativity. While we here focus on brainstorming, the mentioned group effects occur in other types of (human) teamwork as well; thus, future research should explore the generalizability of this lens.

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

13.7 References

- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.

- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Dugosh, K. L., Paulus, P. B., Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *J. Pers. Soc. Psychol.*, 79(5), 722–735. <https://doi.org/10.1037/0022-3514.79.5.722>
- Gozzo, M., Woldendorp, M. K., & Rooij, A. de. (2022). Creative Collaboration with the “Brain” of a Search Engine: Effects on Cognitive Stimulation and Evaluation Apprehension. In M. Wölfel, J. Bernhardt, & S. Thiel (Eds.), *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. ArtsIT, Interactivity and Game Creation* (Vol. 422, pp. 209–223). Springer. https://doi.org/10.1007/978-3-030-95531-1_15
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Hwang, A. H.-C., & Won, A. S. (2021). IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *CHI '21 Proceedings*. ACM. <https://doi.org/10.1145/3411764.3445270>
- Ivanov, A., & Cyr, D. (2014). Satisfaction with outcome and process from web-based meetings for idea generation and selection: The roles of instrumentality, enjoyment, and interface design. *Telematics and Informatics*, 31(4), 543–558. <https://doi.org/10.1016/j.tele.2013.12.004>
- Lambe, K. A., O'Reilly, G., Kelly, B. D., & Curristan, S. (2016). Dual-process cognitive interventions to enhance diagnostic reasoning: A systematic review. *BMJ Quality & Safety*, 25(10), 808–820. <https://doi.org/10.1136/bmjqs-2015-004417>
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.*, 37(6), 822–832. <https://doi.org/10.1037/0022-3514.37.6.822>
- Liu, M., Ke, W., & Xu, D. J. (2023). Will Humans be Free-Riders? The Effects of Expectations for AI on Human-AI Team Performance. *PACIS 2023 Proceedings*.
- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429

- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.
- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Price, K. H., Harrison, D. A., & Gavin, J. H. (2006). Withholding inputs in team contexts: Member composition, interaction processes, evaluation structure, and social loafing. *Journal of Applied Psychology*, 91(6), 1375–1384. <https://doi.org/10.1037/0021-9010.91.6.1375>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.
- Siemon, D., & Wank, F. (2021). Collaboration With AI-Based Teammates – Evaluation of the Social Loafing Effect. In *PACIS 2021 Proceedings*. <https://aisel.aisnet.org/pacis2021/146/>
- Stieglitz, S., Mirbabaie, M., Möllmann, N. R. J., & Rzyski, J. (2022). Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. *Information Systems Frontiers*, 24(3), 745–770. <https://doi.org/10.1007/s10796-021-10201-0>

14 Publication 5: Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration

Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>

Abstract

With the widespread adoption of generative large language models (GLMs) such as GPT-3 or ChatGPT for human-AI problem solving, understanding the effect on performance becomes important. Brainstorming is an established approach for generating ideas to solve problems. In this study, we investigate how AI ideas affect the brainstorming performance metric ‘flexibility’, which refers to the breadth of exploration or coverage of the topic. The foundation for our analysis is the data from an experiment (n=52) in which individual participants brainstormed in two conditions: (1) human-only (baseline) and (2) human+AI (treatment). The treatment condition had access to ideas generated via the GLM OpenAI GPT-3.5. Results show significantly higher flexibility for the human+AI as compared to the human-only condition with a large effect size. With our study, we contribute to the literature of electronic brainstorming, brainstorming with GLMs, as well as to the research challenge of human-AI collaboration.

Keywords: Brainstorming, Human-AI Collaboration, GPT-3.5, Generative Language Models, Performance

14.1 Introduction

Brainstorming is a popular technique for groups of humans to generate ideas for solving problems (Osborn, 1953). However, other humans are not always cost-effectively available. While using technology to facilitate human brainstorming groups is common, tools to support humans with actual ideas (e.g., Siemon et al., 2015) are scarce. With recent advancements of generative large language models (GLMs), however, new opportunities might arise. GLMs were used for creative tasks (Gero et al., 2022), both in free form, such as in the use of ChatGPT, as well as in embedded form with a use case specific graphical user interface. More specifically, several studies successfully used GLMs such as ChatGPT, GPT-3, or GPT-4 for generating ideas (Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023), even suggesting that GLMs' creative abilities might be comparable to human creative abilities (Haase & Hanel, 2023). Thus, in this study, we go beyond considering the technical system merely as a tool as well as the human and the technical system separately but instead explore human-AI collaboration. In an approach similar to Di Fede et al. (2022), we set up a GLM-based brainstorming app, allowing users to request idea suggestions from a GLM for the brainstorming question. The GLM might

thus benefit the human by contributing ideas similar to a human (Haase & Hanel, 2023) or by providing inspiration.

However, such GLMs have limitations. With many of them being trained on large datasets from the internet, GLMs may reproduce bias from the training data and show limited output diversity (Bender et al., 2021; Lin et al., 2022). This calls into question how effectively such systems can be used in a creative brainstorming setting, for which one outcome measure frequently used is flexibility, i.e., breadth of exploration or coverage of a topic (Althuizen & Reichel, 2016; Nijstad et al., 2010). We therefore seek to answer the following research question (RQ):

How does generating ideas jointly with a generative language model affect the breadth of exploration (flexibility)?

To answer this research question, we used data collected for a previous study (under review) focusing on human-AI brainstorming. We developed a prototype for a GLM-based brainstorming app (Figure 1). We used the powerful OpenAI GPT-3.5 model (OpenAI) as a technical foundation. We had participants brainstorm for 10 minutes on a societal problem. Participants were assigned to either the human-only (baseline) or the human+AI (treatment) condition, enabling a quantitative group comparison. While the previous study focused on comparing the outcomes regarding quantity, novelty, and value of individual ideas, the study at hand focuses on the breadth of exploration and, thereby, on assessing the quality of sets of ideas.

Figure 1. GLM-based brainstorming app prototype for data collection

To measure flexibility, we followed the well-established approach of classifying ideas into (pre-defined) categories (Althuizen & Reichel, 2016; Nijstad et al., 2010; Ritter & Mostert, 2018). Counting the number of categories covered by the ideas generated by the participants provides a measure of how broadly the problem was addressed. We statistically performed a group comparison on the breadth of coverage depending on the condition. Additionally, we perform a more in-depth analysis on the breadth of ideas proposed by the AI and assess who drives the exploration of new categories. We find the human+AI team covers the brainstorming question significantly more broadly as compared to the humans who work alone.

With our work, we contribute to the long research history of brainstorming research (Osborn, 1953), more specifically, using technology to enhance brainstorming performance (Pinsonneault et al., 1999). With our tool, we move beyond merely facilitating brainstorming (on a meta-level) but examine a system that supports through actual ideas similar to a human (content-level).

Additionally, we contribute to the emergent literature of exploring challenges, limitations, or boundaries when using GLMs (Floridi & Chiriatti, 2020; Lin et al., 2022). We find that while GLMs were shown to reproduce bias, the output on our societal problem was diverse with regard to the categories covered. We thereby sharpen the understanding of how GLMs may support work-related idea generation.

More broadly, we contribute to the emergent challenge of problem-solving through human-AI collaboration (Akata et al., 2020; Dellermann et al., 2019; Krogh, 2018) and improving the understanding of team dynamics of humans working alongside AI (Makarius et al., 2020), showing how turning individual work of a human into collaborative work of human and AI affects an important brainstorming metric.

14.2 Background

14.2.1 Brainstorming

Brainstorming is a popular approach for groups of humans to generate ideas for solving problems (Osborn, 1953). For effective brainstorming, Osborn (1953) suggested four rules to follow: (1) delayed judgment, (2) encouragement of ss, (3) quantity of ideas, and (4) combining and improving each other's ideas. Since then, lots of research focused on understanding and improving brainstorming performance.

Brainstorming performance can be measured in a variety of ways. Typical measures include the quantity ("fluency"; Nijstad et al., 2010) and quality of ideas produced. Quality can be assessed on the level of individual ideas, e.g., considering the novelty or value of ideas (previous study). Additionally, quality can be assessed on the level of sets of ideas, which we will focus on in this study. An important measure is how broadly the brainstorming question is covered (Althuizen & Reichel, 2016; Nijstad et al., 2010). A typical way to conceptualize coverage is by imagining there to be different aspects or categories of ideas for a brainstorming question. A typical way for assessing the breadth of coverage is by explicitly developing a category system for ideas to reflect the different aspects of a topic and sorting all ideas of the brainstorming session into these developed categories. For each brainstorming session, the breadth can then be calculated as the number of distinct categories covered in this session, from zero (no idea) to the number of categories within the category system. The category system can be developed inductively based on the ideas (e.g., Althuizen & Reichel, 2016) or adapted from prior research.

Past decades of research surfaced many performance-enhancing and -reducing brainstorming effects (Pinsonneault et al., 1999). While Pinsonneault et al. (1999) list 16, we here describe the two most relevant effects for our research question: cognitive stimulation (performance-enhancing) and cognitive inertia (performance-reducing).

Cognitive stimulation refers to the effect that the “utterance of [brainstorming group] members may contain task related stimuli that elicit new ideas from other members” (Pinsonneault et al., 1999). Providing stimuli, e.g., through a confederate or played via audio tape, can improve brainstorming performance (Dugosh et al., 2000; Paulus et al., 2013). However, cognitive stimulation only occurs if humans pay attention to the stimuli (Leggett Dugosh & Paulus, 2005). Additionally, stimuli can have different effects on the way the problem or solution space is explored, with conceptually more distant stimuli encouraging a broader exploration, whereas more closely related stimuli encourage a more in-depth exploration (Althuizen & Reichel, 2016; Althuizen & Wierenga, 2014).

Cognitive inertia, on the other hand, refers to group members “embark[ing] on a single train of thought, which limits creativity and productivity” (Pinsonneault et al., 1999). Typically, cognitive inertia is thought to occur in *nominal groups*, i.e., when group members brainstorm individually, and their ideas are pooled afterward (as compared to brainstorming together as part of *real groups*). In nominal settings, group members do not benefit from outside stimuli and might produce more ideas that are similar, i.e., exploiting existing categories instead of discovering new categories, resulting in a higher (average) number of ideas in only a few categories (higher *within-category fluency*; Nijstad et al., 2010). We discuss both effects with regard to our experiment in the next section.

14.2.2 Human-AI brainstorming

Besides using AI for decision-making, Krogh (2018) suggests exploring using AI for solving problems. Indeed, more recently, combining humans and AI systems to solve problems was discussed as a new research challenge (Akata et al., 2020; Dellermann et al., 2019). There is a long research history of trying to make brainstorming more effective using technology – researched under the label of electronic brainstorming (Pinsonneault et al., 1999). Such technological support can take the form of facilitation, e.g., through processual guidance or through appropriate design decisions to encourage desired behavior.

Besides such meta-level support, tools contributing on content-level have been explored. Such tools can provide stimuli like words, or partial or full ideas related to the brainstorming question. Such systems might use existing content from social media (Siemon et al., 2015) or curated association dictionaries (Althuizen & Reichel, 2016).

Using AI, however, might be difficult, as beforehand, it is unclear which brainstorming questions the user will want to use the tool for, making gathering training data and training an AI difficult. However, with advances of AI, particularly GLMs, new opportunities arise. GLMs such as GPT-3, GPT-4, or ChatGPT (OpenAI) are trained on a large corpus of text

to complete the next word given a certain input. Such systems have shown remarkable results on traditional natural language processing tasks (Brown et al., 2020) as well as on creative tasks (Gero et al., 2022; Q. Zhu & Luo, 2022). Such systems are pre-trained, requiring no task-specific training or fine-tuning (Brown et al., 2020). More specifically, their creative potential was explored for brainstorming (Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023).

Going beyond investigating the technical system and the human separately, Di Fede et al. (2022) proposed to leverage GPT-3 to support humans in brainstorming. After demonstrating the feasibility of using GPT-3 in such human-AI brainstorming settings (Memmert & Tavanapour, 2023), we now quantitatively investigate this setting.

One important goal for human-AI collaboration is the *superior performance* of the human-AI team as compared to the individual (Dellermann et al., 2019). We thus propose to investigate how turning individual work of humans into collaborative work of humans and AI systems affects performance. Besides the direct effect of the GLM proposing ideas, adding a GLM might also affect the human's idea generation. Earlier, the two effects (*cognitive stimulation* and *cognitive inertia*) were introduced. These effects are known from all-human groups. Given that humans are known to respond socially to technical systems (Nass & Moon, 2000) and GLMs' creative abilities were described to be comparable to humans' creative abilities (Haase & Hanel, 2023), we suggest applying this lens to our collaborative human-AI setting. However, it is unclear if those effects occur in such an interactive human-AI setting and how they affect performance overall.

As part of this study, we provide participants access to ideas for a brainstorming question, which are generated by a GLM. Such suggestions could act as stimuli, which might lead to cognitive stimulation. In a previous study (Memmert & Tavanapour, 2023), participants reported having felt inspired to explore new areas of the problem (i.e., *cognitive stimulation*). This might lead to the assumption that access to such AI suggestions could increase flexibility. However, other participants reported that the AI suggestions were basically their previous ideas rephrased in other words. In this case, the AI ideas (i.e., stimuli) might not be capable of preventing *cognitive inertia*, typically observed when working alone. On the contrary, showing potential examples of solutions might lead to *fixation* (Lamm & Trommsdorff, 1973; Sio et al., 2015).

As discussed before, the conceptual distance of stimuli affects how broadly or deeply the problem and solution spaces are explored (Althuizen & Reichel, 2016). The nature of GLM's suggestions with regard to conceptual distance, however, is unclear. Perhaps more importantly, there are discussions around GLMs' output quality, e.g., GLMs were shown to reproduce biases, stereotypes, and falsehoods, with limited diversity in outputs (Bender et al., 2021; Floridi & Chiriatti, 2020; Lin et al., 2022). Particularly, the lack of diversity (Bender et al., 2021) might lead to a narrow exploration, which could result in reduced brainstorming flexibility.

Following the brainstorming rules, in our experiment, the ideas of the human are provided as input to the AI, and humans will have access to the AI ideas. As discussed, the implications on flexibility are unclear, both on the level of the human and on the team overall (human+AI). In previous research on brainstorming with GPT-3, participants subjectively reported signs of both a broadening and narrowing of the perspective due to the AI (Memmert & Tavanapour, 2023); additionally, the conceptual distance and diversity of GLM outputs are unclear. We thus propose the following two undirected hypotheses:

H1. Presence of AI support will affect the diversity of ideas on team level.

H2. Presence of AI support will affect the diversity of ideas on individual level.

While many comparisons might be feasible, we compare performance depending on the presence of AI ideas, resulting in the two conditions: human-only vs. human+AI. A human-AI team achieving superior performance as compared to the individual human is a core ambition of human-AI collaboration research (Dellermann et al., 2019). Additionally, from a practical perspective, other humans might not always be cost-effectively available, which is why we did not include a comparison to human teams at this stage.

14.3 Method

14.3.1 Procedure and participants

For data collection, we developed a GPT-based brainstorming app (see Figure 1 for a screenshot and the next subsection for a more detailed description), similar to the app proposed by Di Fede et al. (2022). We had participants brainstorm individually for 10 minutes on a societal problem. Societal problems are common for such brainstorming studies (Huber et al., 2019). We selected the problem of food waste reduction ('How can we reduce food waste?'), which was used in previous brainstorming studies (Y. Zhu et al., 2021). Afterward, we had participants select their best ideas and answer a post-questionnaire. Participating in the study took about 30 minutes. We conducted the study in three courses taught at our university's informatics department on (1) design science research (two sessions), (2) computer-supported cooperative work, and (3) data-driven solutions for the smart city.

For our experiment, we had two conditions: a baseline condition (human-only), in which each human brainstormed individually, and a treatment condition (human+AI), in which each human brainstormed together with a GLM. As we sought to understand how the GLM affects the results, we asked participants to request suggestions at least once. We communicated to the participants that only suggestions in the "list of ideas" (left side of the screen) would count toward their results (i.e., not-accepted AI ideas are excluded).

14.3.2 Data collection instrument

We collected the data for a different study (under review) with a focus on understanding how AI suggestions affect idea quantity and quality. To collect the data, we developed a

brainstorming app containing the study details, the brainstorming question, and a timer. Participants could add, edit, and remove ideas. Participants in the human+AI condition could request AI suggestions, which were displayed next to their own ideas in a separate list (3 ideas at a time). We decided to allow copying AI suggestions into the participants' "list of ideas" instead of only using suggestions as stimuli to make the scenario more realistic. However, the GLM might also fulfill the role of a stimuli provider.

```

You are part of a brainstorming team. Your goal is to come up with novel
and valuable ideas for the following question: {BRAINSTORMING_QUESTION}
Discarded ideas so far:
{list_ai_ideas}
Novel and valuable ideas so far:
{long_list_ideas}
Please come up with {NUMBER_OF_IDEAS} additional novel and valuable ideas
for the question: "{BRAINSTORMING_QUESTION}". Please provide the {NUM-
BER_OF_IDEAS} additional ideas as enumerated, ordered list. Each idea
should be limited to a maximum of 20 words.

```

Figure 2. Prompt template

Note: curly brackets indicate placeholders; all caps placeholder labels indicate parameters fixed for this study; small caps placeholder labels indicate parameters changing based on the current canvas state; portions in italic font are only included if ideas are already present in the respective lists

To produce the AI suggestions, we used OpenAI's GLM solution 'gpt-3.5-turbo-0301' at a temperature of 0.9, as recommended for creative applications (OpenAI, 2023). We defined a prompt template (see Figure 2) and dynamically populated it with the study details (i.e., brainstorming question) and the current state of the brainstorming pane, i.e., content entered by the participant (ideas) and past AI ideas. The human ideas are provided to the GLM, as an important aspect of brainstorming in groups is to build on each other's ideas (Osborn, 1953). Such reciprocity is also a core aspect of collaboration (Bedwell et al., 2012).

We decided to make users actively request suggestions instead of pushing suggestions proactively, as this was shown to be most effective (Siangliulue et al., 2015). All tool interactions are logged for analysis.

14.3.3 Data preparation and analysis

To investigate the diversity of ideas, we followed the established approach of sorting all ideas into categories (Althuizen & Reichel, 2016; Nijstad et al., 2010; Ritter & Mostert, 2018). After reviewing multiple category systems on 'food waste', we decided to use the system of Specht and Buck (2019). The system consists of 13 categories within five clusters. We selected this category system as it was developed on user-generated contributions (tweets), which are similar in length to the ideas in our study, based on contributions from users from a Western country (US; our participants attend a Western European university), and had an appropriate granularity, i.e., number of categories (compare, e.g., Althuizen & Reichel, 2016). We had a blind-to-condition student assistant sort the ideas into the predefined categories (see Table 1). All but seven ideas were categorized accordingly. We excluded all non-categorized and all deleted ideas from the analysis.

Table 1. System of categories (Specht & Buck, 2019) and exemplary ideas from brainstorming sessions

Area	Category	Exemplary ideas from brainstorming sessions
Domestic or household behavior change	Meal planning	Promote meal planning and portion control to reduce overbuying and food waste at home
	Waste mitigation	Consuming leftover products instead of buying new products
	Smart technology	Utilize technology to create smart refrigerators that track food expiration dates and provide recipe suggestions using expiring ingredients
Food waste diversion and donation	Large-scale food donation	Create a food donation program for excess food from commercial kitchens and events
	Food waste markets	Implement "ugly produce" programs that sell visually imperfect but still edible fruits and vegetables at a discount
Recycling and upcycling	Value-added products	Develop biodegradable packaging made from food waste materials
	Converting food waste into energy	Create a network to distribute food waste from grocery stores and restaurants to biogas facilities for renewable energy production
	Food waste for agricultural purposes	Utilize food surplus for sustainable animal feed to reduce waste in the agriculture industry
Consumer education	Public information campaigns	Create a national awareness campaign that educates consumers on the environmental and social impact of food waste
	Mobile technology	Develop an app that connects individuals with nearby food businesses to donate excess food before it spoils
	Family and consumer science training	Create a food preservation education program to teach individuals how to properly store and preserve food
Governmental action and policy	Legislating food waste reduction	Encourage food donation by providing tax incentives for businesses that donate excess food
	Food date labeling	Implement a food packaging labeling system that indicates the actual shelf life instead of a standardized date

For data analysis, we processed the log data with Python scripts. To calculate descriptive and inferential statistics (incl. assumption checks), we used JASP (JASP, 2023). For visualizations, we used Tableau.

14.4 Results

Our study had 54 participants. We had to exclude the data of two students (due to incomplete data and due to a misunderstanding with regard to the task). The remaining 52 participants (age: mean=22.3, min=18, max=37 years old; gender: 8 female, 44 male) were university students enrolled mostly in informatics study programs. Participants were randomly assigned to conditions and were equally distributed across conditions (human-only: 26, human+ai: 26). In total, 600 ideas were included in the analysis. The distribution across

conditions and idea origin is shown in Table 2. Participants in the human+ai condition requested AI input 148 times (mean=5.7), but only a subset of these ideas is reflected in the final idea set.

Table 2. Number of ideas (and mean values) by origin and condition

Origin Condition	AI	human	Total
human+AI	373 (mean=14.3)	87 (mean=3.3)	460
human-only	-	140 (mean=5.4)	140
Total	373	227	600

14.4.1 Breadth of exploration

We investigate our first hypothesis, exploring whether ‘superior performance’, an important aspect of human-AI collaboration, was achieved in our setting. Superior performance here refers to whether the human-AI team performed better than the human individually, with respect to the metric ‘flexibility’.

We calculate the breadth of coverage for each participant (including the 115 AI suggestions that were accepted/copied by the participant for the human+ai condition into their “list of ideas”) by counting how many categories of the 13 categories (i.e., distinct) each participant covered (mean=4.115). We then perform a group comparison (see Figure 3). As the normality assumption was not fulfilled for the human-only condition (Shapiro-Wilk: $W=.874$, $p=.004$; significant results suggest a deviation from normality), we conducted the non-parametric alternative to the Student’s t-test, the Mann-Whitney U test. The test showed a significant difference ($U=546.5$, $p<.001$) between the human+AI condition ($mdn=5$) and the human-only condition ($mdn=2$), with a strong effect size (rank-biserial correlation $r_B=.617$). Thereby, the first hypothesis is confirmed, and the important goal of ‘superior performance’ of the human-AI team is met.

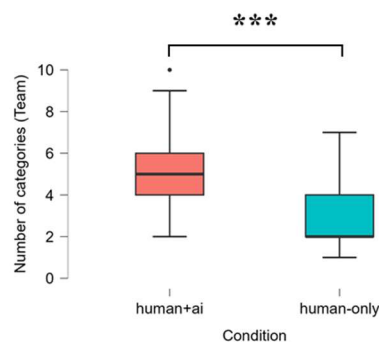


Figure 3. Group comparison on teams’ number of categories covered ($p < .001$)**

We then investigated our second hypothesis, i.e., that humans themselves (i.e., excluding ideas of AI origin in the human+ai condition) cover the question more broadly as compared to the humans working alone (see Figure 4). For each participant, we calculated the breadth of coverage (excluding AI ideas for the human+ai condition) for the 13 categories

(mean=2.846). As the assumption of normality was fulfilled for neither of the conditions (Shapiro-Wilk: human-only: $W=.874$, $p=.004$, human+AI: $W=.902$, $p=.018$), we calculated the non-parametric alternative to the Student's t-test, the Mann-Whitney U test. We found no significant difference between the two conditions ($U=326.0$, $p=.830$). We thus reject our second hypothesis; we do not observe a difference in flexibility for the human with AI ideas present.

A potential explanation could be the number of ideas contributed by the humans. In our previous study (under review), we found humans in the human-only condition to contribute significantly more ideas as compared to the humans in the human+AI condition. Assuming that there might be a tendency to explore more categories the more ideas are contributed, we examined whether a correlation was present. As the normality assumption did not hold, we calculated Spearman's rho. We find a strong, significant, positive correlation (Spearman's $\rho=.817$, $p<.001$), confirming earlier findings of Althuisen and Reichel (2016). Thus, fewer categories explored by the humans in the human+AI condition might be partially attributed to humans contributing fewer ideas in this condition.

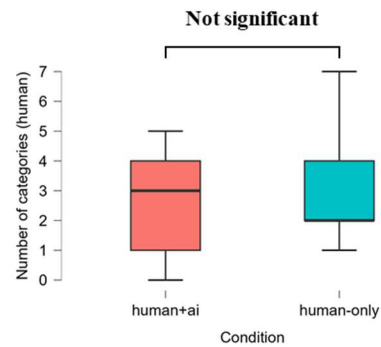


Figure 4. Group comparison on humans' number of categories covered

Humans in the human-only condition contributed significantly more ideas compared to humans in the human+AI condition. However, they did not differ significantly in the number of categories covered. Thus, we expected that the humans in the human-only condition contributed more ideas per category on average (i.e., higher within-category fluency). To test the assumption, we calculated the average number of ideas per category covered by each participant (for participants with no ideas we set 0). The Mann-Whitney U test (assumptions of normality not fulfilled; Shapiro-Wilk: human-only: $W=.771$, $p<.001$, human+AI: $W=.831$, $p<.001$) shows a significant difference ($U=119.0$, $p<.001$) in the average number of ideas per category between human-only ($mdn=2$) and human+ai ($mdn=1$) condition. Thus, humans in the human-only condition show higher within-category fluency.

14.4.2 Driving category exploration

For more context, we performed an analysis on who drove category exploration. To do so, we analyzed, on a summative level, how broadly the categories were explored by both humans and the AI (including 373 AI suggestions). Additionally, we examined who drove the exploration of new categories in the sessions.

For the first analysis, we calculated both for the ideas originating from humans and from the AI how these ideas were distributed across the categories (see Figure 5). As the number of total ideas differ, we show the relative distribution. We find that the AI covers all 13 categories, whereas the humans cover only 12 categories, not providing ideas for the category of ‘value-added products’. Across conditions, most ideas fell into the ‘waste mitigation’ category, whereas fewest fell into the ‘value-added products’ category.

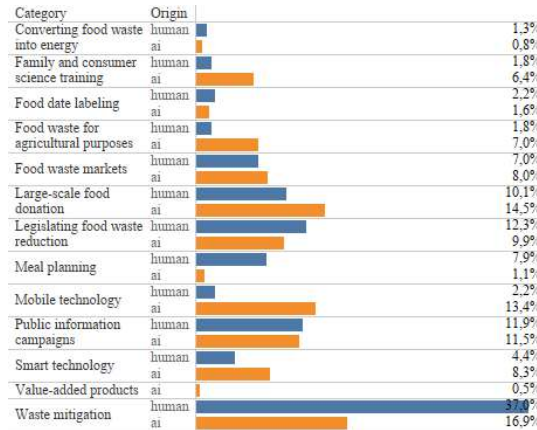


Figure 5. Percentage of ideas across categories of humans (blue) and AI (orange)

On a category-level, we observe the largest difference between human and AI for ‘waste mitigation’. Whereas 37.0% of human ideas fell into this category, it was only 16.9% of AI ideas. The AI put a relatively large emphasis (13.4%) on ‘mobile technology’ compared to the humans (2.2 %). However, these values are not independent, i.e., the humans and the AI system had access to each other’s ideas, as suggested per the brainstorming rules (Osborn, 1953).

We thus additionally take a process perspective, asking who proposes ideas of new categories when the brainstorming counterpart has already contributed ideas. This analysis can only be carried out for the human+AI condition. For the analysis, we order the contributions of both humans and AI systems chronologically within each brainstorming session (per participant). We then examined all cases where the human contributed after the AI made a contribution and vice versa. We find that the AI explores new categories in 41.2% of the cases (i.e., when proposing a new idea), whereas the human explores a new category in 41.9% of the cases.

14.4.3 Subjective perception

To get a more complete understanding of the human+AI collaboration, we asked participants open-ended questions about their experiences. Several participants reported to have felt that the AI helped them explore the topic more quickly and comprehensively:

- “AI can help you to bring more ideas to light” (P29)
- “I was working much faster. Thanks to the AI more ideas came to my mind in less time which made the Brainstorming process much easier than usually” (P28)

Additionally, some participants reported that the AI suggestions helped them to get a new perspective topic, hinting at cognitive stimulation to have occurred:

- “Because of the AI I first came up with certain ideas” (P29)
- “The new ideas of the AI helped me thinking of additional ideas.” (P27)
- “The ideas on the side which came from the AI were a great inspiration for developing new ideas on my own.” (P28)

However, participants also reported that the AI influenced and potentially narrowed their way of thinking, particularly due to repetitive suggestions, which could hint at cognitive inertia to have occurred:

- “[The AI] would often propose very similar ideas” (P37)
- “It felt a lot faster, but a bit repetitive also since the AI started generating ideas similar to the previous ones” and “[...] It definitely felt like my thought process was being governed by the AI [...]” (P52)
- “The AI stopped coming up with original ideas after a few were generated and I couldn't concentrate on coming up with my own ideas” (P34)
- “[...] set my focus in the direction of the suggestions” (P42)

Overall, the feedback fits the quantitative team-level performance results, with many participants stating that the AI helped them to cover the topic more comprehensively. For individual performance, while some participants reported having gained a new perspective, others responded negatively to the AI, with one stating “Maybe [it] cut[...] off my creativity” (P33).

14.5 Discussion

14.5.1 Answer to the research question

Flexibility, i.e., breadth of exploration or coverage, is an important brainstorming performance measure, as it is strongly correlated with a high number of high-quality ideas (Nijstad et al., 2010). With the increasing adoption of GLMs (e.g., GPT-3, ChatGPT) for creative tasks in general and generating ideas more specifically, it becomes important to understand how GLMs affect how broadly or narrowly the problem and solution space is explored. Particularly so, as creativity is typically considered a human strength (Dellermann et al., 2019), and AI systems are known to potentially lack diversity in outputs (Bender et al., 2021).

For our setting, we find that the human working with the AI jointly produced significantly more ideas as compared to alone. Both human and AI seem to have similarly driven the exploration of new categories. This is interesting, given the assumed superiority of humans over AI in creative tasks, but is in line with more recent findings of Haase and Hanel (2023),

who found GLMs to have “comparable to human creative abilities” in certain aspects. Furthermore, the AI produced ideas more evenly across the categories. Thus, the brainstorming performance concerning flexibility seems to be affected positively by the presence of the AI.

However, adding the AI did not lead to more ideas on the level of the human. Other than one might expect according to *cognitive stimulation*, humans did not produce more ideas or cover more categories with their ideas. For the former, one reason could be that the participants working with the AI spend time reviewing and selecting ideas, reducing the time available for brainstorming. Such a phenomenon was already described by Pinsonneault et al. (1999, p. 126) and coined the “distraction effect”, explained as “individuals [...] spending too much time reading others’ ideas rather than thinking about new ideas, thus inadvertently limiting their productivity”. This could be a potential (partial) explanation as to why no significant difference was observed with regard to the number of ideas for the humans between the conditions. Given that the number of ideas is correlated with the number of categories covered, this might then (partially) explain the lack of a significant difference in the categories covered. However, future research is required on this aspect.

14.5.2 Contribution and implications

Our work provides theoretical contributions to the literature streams of brainstorming with GLMs, brainstorming group effects as well as human-AI collaboration more broadly and offers practical contributions for supporting brainstorming with GLMs.

We contribute to the discourse around GLMs and brainstorming (Haase & Hanel, 2023; Koivisto & Grassini, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023), enhancing the understanding of the potential of using GLMs for creative idea generation. Other than previous work, we do not investigate humans and GLMs separately but offer insights into joined human-AI brainstorming sessions as suggested, e.g., by Di Fede et al. (2022). Our analysis aligns with and expands on the findings of Haase and Hanel (2023). Not only do GLMs produce ideas of comparable novelty (Haase & Hanel, 2023), but we show that GLMs also drive category exploration comparable to humans.

With our work, we contribute to the literature of electronic brainstorming and related group effects (Pinsonneault et al., 1999). Due to moving beyond an isolated setting in which humans and GLMs work separately, we offer insights into the applicability of known group effects (*cognitive stimulation*, *cognitive inertia*) from all-human groups to human-AI groups. We find that participants’ reports contain signs of cognitive stimulation, with some stating they explored new areas due to the AI. Thus, GLMs might be capable of taking the role of a stimuli provider. However, we did not observe a broader exploration of the topic by the human. We offer a potential explanation, the distraction effect, which could offset the performance-enhancing cognitive stimulation effect, but future research should explore how this perceived stimulation materializes.

Participant's reports also contained signs of cognitive inertia, with participants explaining that the AI kept them on their train of thought, which could mean the AI-induced fixation (see Lamm & Trommsdorff, 1973). Quantitatively, the AI does not seem to affect flexibility of the individual humans when measured by the breadth of coverage, i.e., the number of categories covered. However, we did observe participants in the human-only condition contributing more ideas within the categories covered (higher within-category fluency), which could hint at them not leaving their "train of thought", i.e., cognitive inertia (Lamm & Trommsdorff, 1973; Pinsonneault et al., 1999). A potential interpretation could be that the GLM led humans to not focus on only a few categories but to add ideas within other (AI-explored) categories. We thus call for more in-depth research to reconcile these observations.

More broadly, in taking a collaborative perspective, we contribute to the research challenge of solving problems through human-AI collaboration (Akata et al., 2020; Dellermann et al., 2019). We demonstrate a core ambition of human-AI collaboration (superior performance) regarding a key brainstorming metric (flexibility). We show the importance of the GLM in driving this performance, which is remarkable, given the traditionally assumed human superiority in creative tasks (Dellermann et al., 2019).

We also offer a practical contribution to brainstorming by describing an instantiated information system supporting humans to develop ideas for a problem. We demonstrate that even without further training data or fine-tuning and without relying on copying existing content from other platforms or on the manual preparation of stimuli (Althuizen & Reichel, 2016; Siemon et al., 2015), modern GLMs can enhance the creative ability of humans when forming a human-AI team. Given our results, one might encourage using GLMs for brainstorming (particularly in the early phases), as this enables humans to more broadly explore the question, which is known to be correlated with a larger number of high-quality ideas. Future research should explore individual differences (e.g., regarding creative ability or topic knowledge).

14.5.3 Limitations

Our study has several limitations. We only covered one brainstorming question (i.e., food waste), limiting generalizability. However, using societal problems for brainstorming studies is common, and the specific question was used in brainstorming research before (Y. Zhu et al., 2021). We did not adjust the AI system regarding this specific brainstorming question; on the contrary, the question can simply be replaced during the prompt template population (see Figure 2).

The results of our study are dependent on how flexibility is measured. We selected a pre-defined, published category system to increase objectivity. We reported our reasoning for selecting this category system. However, other category systems exist, which might have led to different results. Closely connected to this point: while we did cover idea flexibility, we did not investigate the cultural diversity of suggestions. Some of the ideas might not be

representative of all cultures. The category system was based on the tweets from the users from the US; students in our study attended a Western European university). Besides, using a category system is only one approach to assess flexibility; other approaches include, e.g., semantic distances. However, the approach of using a category system is common. Additionally, besides assessing the idea quality on the level of sets of ideas, assessing the effect of using GLMs on the quality of individual ideas is an important area for future research.

On a more technical note, we only used one prompt template. While we used prompt engineering, a different prompt might have yielded different results. The results might be considered a baseline, as we neither optimized for breadth nor depth of exploration. Lastly, we only used one GLM. However, the model is an advancement on a model that showed high performance across several typical and non-typical natural language model tasks (Brown et al., 2020) and the underlying model for the widely adopted OpenAI product ChatGPT (OpenAI), making it highly relevant to practice.

14.6 Conclusion

GLMs become more widely adopted in work settings to solve problems. When leveraging such systems, it is important to understand the implications on the work results. In our study, we show that teams of humans and AI can outperform humans brainstorming individually on the metric of flexibility, i.e., the breadth of exploration or number of idea categories covered. We contribute to literature on electronic brainstorming (e.g., Althuizen & Reichel, 2016; Pinsonneault et al., 1999), GLMs for brainstorming (Di Fede et al., 2022; Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023), and the research challenge of human-AI collaboration for problem-solving (Akata et al., 2020; Dellermann et al., 2019; Makarius et al., 2020).

Acknowledgment

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

14.7 References

- Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8), 18–28.
- Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of Management Information Systems*, 33(1), 11–44. <https://doi.org/10.1080/07421222.2016.1172439>

- Althuizen, N., & Wierenga, B. (2014). Supporting Creative Problem Solving with a Case-Based Reasoning System. *Journal of Management Information Systems*, 31(1), 309–340. <https://doi.org/10.2753/MIS0742-1222310112>
- Bedwell, W. L., Wildman, J. L., DiazGranados, D., Salazar, M., Kramer, W. S., & Salas, E. (2012). Collaboration at work: An integrative multilevel conceptualization. *Human Resource Management Review*, 22(2), 128–145. <https://doi.org/10.1016/j.hrmr.2011.11.007>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. <https://doi.org/10.1145/3442188.3445922>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Dugosh, K. L., Paulus, P. B. [P. B.], Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *J. Pers. Soc. Psychol.*, 79(5), 722–735. <https://doi.org/10.1037/0022-3514.79.5.722>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing using Language Models. In F. ` Mueller, S. Greuter, R. A. Khot, P. Sweetser, & M. Obrist (Eds.), *Designing Interactive Systems Conference* (pp. 1002–1019). ACM. <https://doi.org/10.1145/3532106.3533533>
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Huber, B., Shieber, S., & Gajos, K. Z. (2019). Automatically Analyzing Brainstorming Language Behavior with Meeter. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–17.
- JASP. (2023). (Version 0.17)[Computer software].

- Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1), 13601. <https://doi.org/10.1038/s41598-023-40858-3>
- Krogh, G. von (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404–409. <https://doi.org/10.5465/amd.2018.0084>
- Lamm, H., & Trommsdorff, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. *European Journal of Social Psychology*, 3(4), 361–388. <https://doi.org/10.1002/ejsp.2420030402>
- Leggett Dugosh, K., & Paulus, P. B. (2005). Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology*, 41(3), 313–320. <https://doi.org/10.1016/j.jesp.2004.05.009>
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. <https://doi.org/10.1016/j.jbusres.2020.07.045>
- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nijstad, B. A., Dreu, C. K. W. de, Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1), 34–77. <https://doi.org/10.1080/10463281003765323>
- OpenAI [Computer software]. www.openai.com
- OpenAI. (2023). <https://beta.openai.com/docs/api-reference/completions/create#completions/create-temperature>
- Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.

- Paulus, P., Kohn, N. W., Arditti, L. E., & Korde, R. M. (2013). Understanding the Group Size Effect in Electronic Brainstorming. *Small Group Research*, 44(3), 332–352. <https://doi.org/10.1177/1046496413479674>
- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Ritter, S. M., & Mostert, N. M. (2018). How to facilitate a brainstorming session: The effect of idea generation techniques and of group brainstorm after individual brainstorm. *Creative Industries Journal*, 11(3), 263–277. <https://doi.org/10.1080/17510694.2018.1523662>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.
- Sio, U. N., Kotovsky, K., & Cagan, J. (2015). Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies*, 39, 70–99. <https://doi.org/10.1016/j.destud.2015.04.004>
- Specht, A. R., & Buck, E. B. (2019). Crowdsourcing Change: An Analysis of Twitter Discourse on Food Waste and Reduction Strategies. *Journal of Applied Communications*, 103(2).
- Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). *Putting GPT-3's Creativity to the (Alternative Uses) Test*.
- Summers-Stay, D., Voss, C. R., & Lukin, S. M. (2023). Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.
- Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, 2, 1825–1834. <https://doi.org/10.1017/pds.2022.185>
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2021). The effect of rank-ordering strategy on creative idea selection performance. *Eur. J. Soc. Psych.*, 51(2), 360–376. <https://doi.org/10.1002/ejsp.2743>

15 Publication 6: Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load

Memmert, L., Cvetkovic, I., Tavanapour, N., & Bittner, E. (2025). Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load. *Business & Information Systems Engineering*. Advance online publication. <https://doi.org/10.1007/s12599-025-00974-y>

Abstract

The adoption of generative AI (artificial intelligence) has sparked increased interest in how individual humans can generate ideas and solve problems collaboratively with AI, aiming to achieve superior performance compared to working alone. Based on group effects from all-human brainstorming groups, the paper investigates a human–AI brainstorming setting using OpenAI’s GPT-3.5 as an embedded generative large language model (GLM). In a between-subjects experiment (n = 75) comparing solitary humans and human–GLM dyads, results show that humans do not perform better individually when supported by a GLM. However, collectively, the human–AI dyad achieves superior (or complementary) performance on common brainstorming performance metrics (fluency, flexibility, novelty, and value). Findings are discussed in relation to effort allocation and potential loafing behavior when working with GLMs. The paper advances our understanding human–AI group dynamics, the transferability of group mechanisms (cognitive stimulation, free riding, cognitive inertia) from all-human to human–AI groups, and the discourse on smart loafing with AI group members.

Keywords: Brainstorming; Human-AI collaboration; GPT; Large language model; Creativity

15.1 Introduction

Generating ideas is the divergent stage of a larger process for solving creative problems in organizations, such as developing new products or marketing slogans (Bouschery et al., 2023; Dell’Acqua et al., 2023; Sawyer, 2021). A popular idea-generation technique is brainstorming (Maaravi et al., 2021). Traditionally, it was designed for human groups. Today, advances in generative AI allow a single person to brainstorm with an AI system rather than with a colleague (Holzner et al., 2025; Kumar et al., 2025; Memmert, Mies, & Bittner, 2024).

As generative language models (GLMs) such as ChatGPT are rapidly adopted and often used for on-demand ideation when human partners are unavailable (Jiang et al., 2024), it becomes both timely and relevant to understand their influence on brainstorming outcomes (Kumar et al., 2025). GLMs, a term we use interchangeably with large language models (LLMs), refers to generative AI systems pre-trained on extensive text corpora that can generate coherent, contextually appropriate text in response to user input. Specifically, we examine how GLMs affect classical performance dimensions — fluency, flexibility, novelty,

and value (Dean et al. 2006; Holzner et al. 2025) — and participants' cognitive load (Gonzalez et al. 2024). Particularly as concerns have been raised that working with AI may invite reduced effort, i.e., loafing-like behavior (Stieglitz et al. 2022; Liu et al. 2023) or, conversely, increase cognitive load, for instance, because humans evaluate the outputs (Tankelevitch et al. 2024), it is vital for tool users and designers as well as organizations to understand GLMs' effects.

Using GLMs for brainstorming extends the long tradition of leveraging technology to support, rather than replace, human creativity. Decades of work on creativity-support systems show how tools can assist individuals during idea generation and, therefore, provide a useful conceptual foundation. Prior research has documented both the metrics used to assess brainstorming performance (Dean et al., 2006) and the mechanisms within human groups that enhance or reduce them (Dugosh et al., 2000; Pinsonneault et al., 1999). Among the most prominent are cognitive stimulation — the additional ideas produced when individuals are exposed to others' ideas — and social loafing — the tendency to reduce effort when working with others because one's contribution seems less essential (Pinsonneault et al. 1999). These two mechanisms form the theoretical backdrop for our investigation of human–GLM brainstorming.

Building on these insights, individual creativity-support systems (ICSS) were developed to help individuals generate ideas more effectively. So-called stimuli-provider ICSS supply brief cues such as single words or short phrases that can trigger cognitive stimulation, but ICSS typically cannot generate complete ideas dynamically (Wang & Nickerson, 2017). We argue that GLMs represent a qualitative leap beyond such systems in three key respects (cf. Brown et al., 2020). First, they are generative rather than retrieval-based: instead of offering predefined cues, they produce original, contextually coherent text by predicting word sequences. Second, they operate across a broad semantic range, drawing on knowledge encoded across vast corpora rather than limited domain databases. Third, they are adaptive: their outputs can be conditioned on the user's prior ideas and requests, enabling iterative exchanges that resemble human-to-human ideation. Together, these properties allow GLMs to produce complete, context-specific ideas in real time, with creative performance approaching that of many humans (Haase & Hanel, 2023; Holzner et al., 2025). Hence, integrating GLMs into brainstorming changes the nature of creative interaction itself — from systems that merely inspire humans to partner-like systems that co-generate ideas.

This leap in generative capability fundamentally reshapes how brainstorming unfolds (e.g., Gonzalez et al., 2024; Houde et al., 2025; Kumar et al., 2025; Specker et al., 2025; Tao et al., 2023; Zhan et al., 2024). In human dyads, two humans actively supply complete ideas that build on each other, whereas stimuli-provider ICSS only (reactively) provide brief cues that the one human must develop further. A human–GLM dyad may lie between these cases: one human actively contributes complete ideas but can also prompt the GLM to

generate complete, copy-ready ideas on demand. Nonetheless, the GLM remains fundamentally reactive. Prior research has mapped interaction dynamics in both human groups (Pinsonneault et al., 1999) and settings involving stimuli-provider ICSS (Gabriel et al., 2016; Wang & Nickerson, 2017, 2019). This reactive but capable role makes the human–GLM dyads distinct from both human-only groups and traditional ICSS setups, underscoring the need for dedicated theoretical and empirical study. At the same time, insights from both prior streams — group brainstorming and ICSS research — help anticipate how GLM-based collaboration might enhance or impede brainstorming performance.

Because a human–GLM dyad includes only one human partner, the individual must continually decide whether to invest effort in crafting ideas personally or to request the GLM’s copy-ready suggestions, and, once received, whether to refine or adopt them. This effort-allocation trade-off may be obscured in multi-human–GLM sessions, where social pressures and peer ideas may also influence contributions (Pinsonneault et al., 1999). By eliminating human-human dynamics, the dyad setting allows us to isolate the effects of high-quality, reactive GLM output on (1) the individual’s idea production and (2) the dyad’s joint output as well as any changes in cognitive load. Following the ICSS research tradition, we therefore compare the dyad with a solitary brainstormer who faces no external ideas or social influence, and, therefore, no comparable trade-off decisions. At the individual level, early studies of GLM-supported brainstorming with individuals or small groups report mixed findings: GLM ideas can make idea generation feel easier (Nomura et al., 2024; Specker et al., 2025; Zhang et al., 2025) but do not necessarily lead humans to produce more or more original ideas (Kumar et al., 2025). Accordingly, our first research question examines how human brainstorming performance differs when considering *only ideas originating from the human* — comparing those who work with a GLM to those who brainstorm alone.

A central ambition of human–AI collaboration is to achieve superior or complementary performance, that is, performance in the human–AI group (here, the dyad) that exceeds what a human can achieve alone (Dellermann et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021). Given GLMs’ ability to generate human-level creative ideas and the potential presence of both positive (e.g., cognitive stimulation, cf. Gonzalez et al., 2024; Specker et al., 2025) and negative group mechanisms (e.g., social loafing, cf. Liu et al., 2023, or the lack thereof, cf. Nomura et al., 2024) in human–GLM brainstorming (cf. Zhan et al., 2024), the overall effect on performance remains uncertain. Studies on individuals or human groups suggest that GLM use can positively affect the number of ideas and topic coverage (Specker et al., 2025), yet it is unclear how these benefits translate to human–GLM dyads. Recent meta-analytic evidence (Holzner et al. 2025) indicates that while generative-AI collaboration modestly enhances creative performance overall, the effects vary substantially across tasks and even different ideation task types. Hence, for the kind of ideation task examined in our study, the impact of human–AI collaboration remains an open question.

This underscores the need for more granular, mechanism-oriented research to better understand GLMs impact in brainstorming contexts. Thus, second, we examine how *dyad-level brainstorming performance* differs from that of human brainstorming alone.

Lastly, it was hypothesized that humans might use AI to maintain or reduce cognitive resources (Jiang et al., 2024; Stieglitz et al., 2022). However, when brainstorming with a GLM, humans must balance their effort between generating their own ideas and leveraging or evaluating the GLM's suggestions (Gonzalez et al., 2024), which might require additional cognitive resources (Tankelevitch et al., 2024). Accordingly, our third research question examines whether working with a GLM influences *perceived cognitive load* as compared to brainstorming alone.

To address these research questions, we conducted a between-subjects experiment ($n = 75$) in which participants brainstormed using a purpose-built app. In the treatment condition, the app allowed participants to request ideas from a GLM on demand, whereas in the control condition (human-only), participants received no GLM support. Brainstorming performance was assessed across four established dimensions — fluency, flexibility, novelty, and value — at both the individual (human) and dyad (human-GLM) levels. After the session, participants completed a cognitive load scale and answered open-ended questions about their experience. Overall, human-GLM dyads performed significantly better than solitary humans. However, contrary to expectations, when considering only ideas originating from humans (excluding GLM ideas), participants in human-GLM dyads did not outperform those working alone and reported no significant difference in perceived cognitive load.

By investigating human-GLM brainstorming, we contribute to the long-standing information systems research stream on ICSS (Althuizen & Reichel, 2016; Pinsonneault et al., 1999; Siemon et al., 2015; Wang & Nickerson, 2017, 2019). We show that GLMs influence humans in ways that differ from those of traditional, stimuli-provider ICSS. Whereas stimuli-provider ICSS seek to improve individual brainstorming performance, e.g., the supported human produces more ideas, GLMs — contrary to expectations — do not improve individual brainstorming performance and may even reduce the number of ideas produced. Moreover, we demonstrate that simply applying the well-documented performance mechanisms from purely human groups (Dugosh et al., 2000; Pinsonneault et al., 1999) to human-GLM dyads can mischaracterize a shift in effort as social loafing. Based on our findings, we offer a nuanced perspective on human effort, reflecting both the new opportunities (e.g., GLMs as capable partners) and challenges (e.g., shift in effort and cognitive load) associated with working with GLMs in creative contexts. In doing so, we contribute towards addressing key research challenges in human-AI collaboration by advancing our understanding of human-AI group dynamics (cf. Benbya et al., 2024; Makarius et al., 2020) and of human-AI superior — or complementary — performance (cf. Dellermann et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021).

15.2 Background

15.2.1 Brainstorming Performance and Determinants

Brainstorming is a popular technique for generating ideas (Maaravi et al., 2021; Osborn, 1953). Osborn (1953) suggested four brainstorming rules: (1) delayed judgment, (2) encouragement of wild ideas and (3) idea quantity, and (4) combining ideas. While initially proposed for groups (Osborn, 1953), different group structures were explored to determine performance factors (Pinsonneault et al., 1999). This included comparing *real groups*, in which members brainstormed together, to *nominal groups*, in which humans brainstormed alone without interaction and without seeing others' ideas (Coskun et al., 2000; Diehl & Stroebe, 1987), i.e., similar to individual or solitary brainstorming.

Common brainstorming performance metrics include the quantity of ideas, so-called *fluency* (Nijstad et al., 2010; Paulus, 2000), and idea quality. For individual ideas, typical quality dimensions are, e.g., *novelty* (or originality), and *value* (or usefulness or feasibility) of ideas (Althuizen & Reichel, 2016; Diehl & Stroebe, 1987; Nijstad et al., 2010; Paulus, 2000). Dimension might combine multiple characteristics, e.g., for rating novelty, the closely related characteristics of originality and surprisingness might be included (Siangliulue, Chan, et al., 2015); hence, we use the terms interchangeably. Ideas are usually rated by blind-to-condition judges (Althuizen & Reichel, 2016) or, more recently, by GLMs (Haase & Hanel, 2023). Considering only the best ideas is common to prevent penalizing idea quantity. This might be achieved by having participants self-select ideas and consider only those shortlisted ideas (Siangliulue, Arnold, et al., 2015). While perhaps realistic for work settings, this introduces uncertainty regarding the selection quality. A different approach is the “good-idea-count” metric (Reinig et al., 2007), in which only ideas rated above a defined threshold are considered. Beyond individual ideas, a key aspect of brainstorming is how narrowly or broadly the topic was covered, i.e., the breadth of exploration or flexibility (Althuizen & Reichel, 2016), which can be assessed by categorizing all ideas of a brainstorming session and counting the number of categories containing at least one idea (Althuizen & Reichel, 2016; Nijstad et al., 2010). The metrics can be calculated on both the individual level and the (collective) group level to understand individual and group performance.

In human brainstorming groups, several group mechanisms rooted in psychological and social theories are known to lead to performance gains or performance losses during the brainstorming session (Pinsonneault et al., 1999). Here, we discuss the three mechanisms of cognitive stimulation, cognitive inertia, and free riding, as these were observed to be relevant for our GLM-based brainstorming setting.

Cognitive stimulation refers to group members creating more or better ideas because they are stimulated by other's ideas; an effect that can occur in real groups when working with others (Pinsonneault et al., 1999) or through creativity support systems (Siemon et al.,

2015), though paying attention to others' ideas is necessary (Dugosh et al., 2000; Leggett Dugosh & Paulus, 2005). Additionally, the number of stimuli and the semantic relatedness of stimuli, i.e., the kind of (categories of) ideas, determine if and how ideation is affected (Baruah & Paulus, 2016; Wang & Nickerson, 2017, 2019).

Cognitive inertia refers to humans “embark[ing] on a single train of thought, which limits creativity and productivity” (Pinsonneault et al., 1999, p. 114), which is typically observed in *nominal groups* (Pinsonneault et al., 1999), where humans brainstorm alone. In contrast to real groups, humans in nominal groups are not exposed to outside stimuli and thus might come up with more ideas in the same category, i.e., higher depth or higher within-category fluency (Nijstad et al., 2010), as opposed to exploring new categories (i.e., higher flexibility).

Free riding or social loafing refers to humans reducing their effort and relying on others to complete the task (Pinsonneault et al., 1999), an effect occurring in real groups (Pinsonneault et al., 1999). Influencing factors include perceived dispensability of effort, i.e., feeling one's contribution would not make much of a difference (Kerr & Bruun, 1983), and diffused responsibility (Latané et al., 1979; Pinsonneault et al., 1999), i.e., feeling less responsible when being part of a group, consequently reducing one's effort (Latané et al., 1979). While those mechanisms were initially discovered in human groups, decades of research explored how to affect these mechanisms through technology (Pinsonneault et al., 1999).

15.2.2 Towards Human–AI Collaboration for Brainstorming

Building on the human group mechanisms above, we examine how these might translate when brainstorming with a tool rather than a human. In human groups, each human contributes ideas directly but can also spark cognitive stimulation through those ideas in the other humans, helping them to overcome inertia (Pinsonneault et al., 1999). However, when humans feel responsibility is shared or their effort is dispensable, e.g., when working with a capable partner, they may reduce their effort, i.e., social loaf (Pinsonneault et al., 1999).

Creativity support systems research investigates tools to enhance brainstorming performance (Wang & Nickerson, 2017), e.g., supporting individuals or groups on a meta-level through procedural facilitation that is brainstorming question independent, or on a content-level, i.e., acting more like a contributor for the specific brainstorming question (e.g., Althuizen & Reichel, 2016; Hwang & Won, 2021; Siemon et al., 2015). In a human-tool dyad, these tools, labelled as stimuli-provider ICSS, supply brief cues like words, phrases, or short sentences related to the brainstorming question. Examples include an “AI-like” system that mines social-media stimuli (Siemon et al., 2015) and a hint generator based on an association dictionary (Althuizen & Reichel, 2016). Such cues act as external stimuli that can spark new ideas and overcome cognitive inertia, much like cognitive stimulation in human groups. Unlike another human, however, ICSS typically do not dynamically and

actively contribute complete ideas; they merely provide inspiration, leaving the human to elaborate the full contribution.

We argue that GLMs mark a qualitative leap. Studies show that text-based generative AI, more specifically, large language models (LLMs) or generative language models (GLMs), can generate creative ideas at humanlike creative levels across a variety of subjects and creative tasks (Haase & Hanel, 2023; Holzner et al., 2025; Organisciak et al., 2023). Replacing an ICSS's backend with a GLM, therefore, enables generating live, content-level support across many topics without prior stimulus curation. Instead of supplying incomplete ideas or cues, GLMs deliver fully formed ideas that can both stimulate the human and can be copied directly to the joint idea pool, much like when brainstorming with another person. Moreover, these systems can dynamically respond to user input, much like a human brainstorming partner could do. Indeed, Gabriel et al. (2016, p. 112) refer to systems capable of suggesting ideas during brainstorming as “colleagues”. While one might argue that GLMs are still just tools and not human, this may fail to acknowledge the new behaviors and dynamics that may emerge in human–GLM dyads (Liu et al., 2023).

Past research suggests that humans might socially respond to technical systems (Nass & Moon, 2000), which could mean that similar group mechanisms arise in human-tool dyads and in human-human dyads (Siemon & Wank, 2021). Indeed, for several ICSS not based on GLMs, the group mechanisms of cognitive stimulation or free riding known from human groups were explored in different constellations with non-humans (Althuizen & Reichel, 2016; Siemon et al., 2015; Siemon & Wank, 2021), including dyads of humans and tools (Althuizen & Reichel, 2016; Stieglitz et al., 2022), and GLMs specifically (Nomura et al., 2024). Recent studies show that humans might attribute responsibility to the AI and might feel their effort to be dispensable (Memmert et al., 2025; Memmert & Tavanapour, 2023; Stieglitz et al., 2022), which are known antecedents for social loafing (Pinsonneault et al., 1999).

In human groups, social loafing is considered a negative behavior, as humans reduce their effort at the expense of other humans. However, when working with AI, loafing might need to be judged differently (Stieglitz et al., 2022). Particularly, if the AI group member performs well, a reduction of effort might be desirable to maintain cognitive resources. Thus, as an adaptation to social loafing, the construct of “smart loafing” was proposed, defined as “the reduction of effort in human-VA collaboration to maintain cognitive resources and enhance efficiency in work” (Stieglitz et al., 2022, p. 758). If the AI group member indeed leads to a reduction in cognitive load, this could be particularly relevant for the creative task of idea generation, because cognitive load and creative performance are negatively associated (Redifer et al., 2021).

Figure 1 summarizes these contrasts in a conceptual framework. Human-human and human–GLM dyads both involve an “other entity” that contributes full ideas, which can also

act as stimulating cues, whereas stimuli-provider ICSS contribute cues only; solitary brainstorming contains neither. Because GLMs share the “full-idea” property with human partners while remaining tools that are triggered by human action (instead of being actively like a human), human–GLM brainstorming research may be informed through the insights from both ICSS and group-brainstorming research, highlighting mechanisms like cognitive stimulation and overcoming inertia, but also cautioning regarding potential loafing behavior.

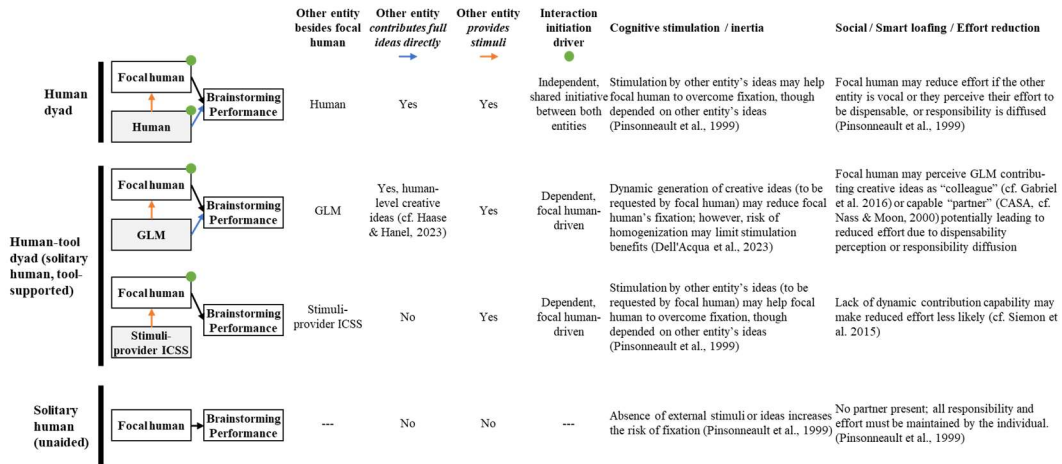


Figure 1 Conceptual framework comparing selected mechanisms in human dyads, human-tool dyads, and solitary humans

Going beyond investigating GLM and human ideation separately (Haase & Hanel, 2023; Organisciak et al., 2023), we explore joint human–GLM sessions in the tradition of content-level ICSS in typical brainstorming studies (Althuizen & Reichel, 2016; Gabriel et al., 2016; Wang & Nickerson, 2017). Though prior research reported a preference for brainstorming with a GLM over brainstorming alone (Houde et al., 2025; Muller et al., 2024), signs for both performance-enhancing and -reducing mechanisms were predicted or observed (Memmert, 2024; Memmert & Bittner, 2024; Specker et al., 2025; Zhan et al., 2024), rendering it uncertain if a GLM-based tool improves performance compared to unaided brainstorming, calling for more research. With our study, we seek to contribute to the discourse on whether and how group mechanisms known from human groups occur when individuals work with AI (Liu et al., 2023; Siemon, 2023; Siemon et al., 2015; Siemon & Wank, 2021; Stieglitz et al., 2022).

15.2.3 Hypotheses Development

We contrast a *human-only baseline*, in which every idea is generated solely by the participant, with a *human+GLM treatment* that lets participants both produce their own ideas and request (and copy) GLM ideas. This conceptual comparison yields two analytically distinct performance layers (Figure 2):

- *Human performance*: ideas authored by the human, and
- *Dyad performance*: the combined output of the human and the GLM.

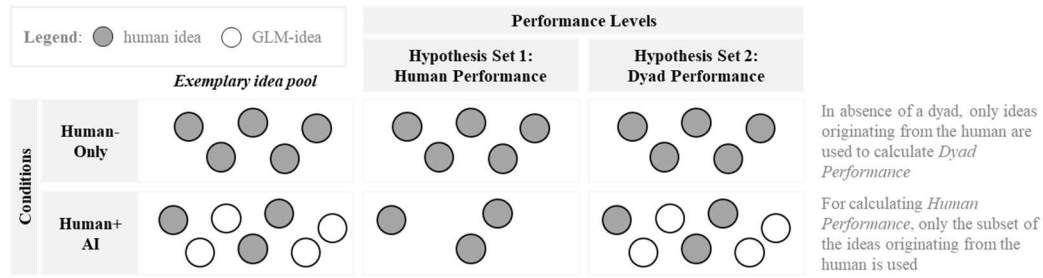


Figure 2 Idea in- and exclusion for hypotheses with performance levels and conditions

Separating these layers is essential for isolating (a) the GLM’s influence on the human’s own contribution and (b) the net benefit of adding a GLM. Accordingly, our hypotheses proceed in three steps. Hypothesis Set 1 tests how a GLM alters individual output; Hypothesis Set 2 asks whether the dyad outperforms the unaided human; and Hypothesis 3 compares self-reported cognitive load. Figure 3 summarizes the specific hypotheses, which we develop next by drawing on established brainstorming mechanisms and recent empirical findings.

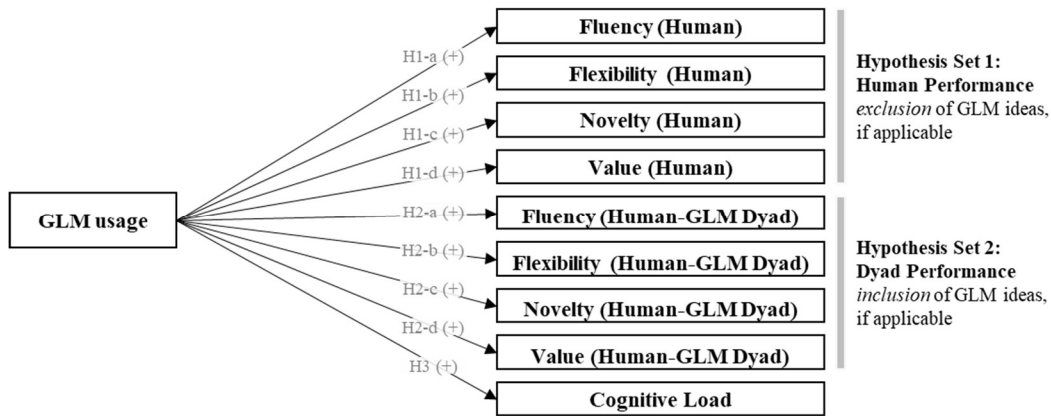


Figure 3 Research Model

15.2.3.1 Hypothesis Set 1: Human Performance

For hypothesis set 1: human performance, we investigate how the *human* brainstorming performance, i.e., considering only ideas originating from the human, excluding GLM ideas, varies across the two conditions, uncovering how typical brainstorming performance dimensions are affected by GLM usage. By adding a GLM, individual human brainstorming is transformed into collaborative human–GLM brainstorming. Previous research on all-human brainstorming groups suggests that individual performance can be enhanced or harmed due to being part of a group (Pinsonneault et al., 1999). While these effects are well-documented in purely human groups, and while humans might socially respond to technical systems (Nass & Moon, 2000), such effects in groups with non-human members (Liu et al., 2023; Siemon & Wank, 2021; Stieglitz et al., 2022; Tao et al., 2023) are still being researched, i.e., it is uncertain whether humans themselves perform differently when working with or without a GLM (Specker et al., 2025; Tao et al., 2023). Previous research suggests group effects like cognitive stimulation, social loafing, or cognitive inertia to be

relevant in our setting (Memmert, 2024; Memmert & Tavanapour, 2023; Stieglitz et al., 2022; Tao et al., 2023).

In brainstorming sessions, human fluency decreases after the first few minutes (Baruah & Paulus, 2016; Kohn & Smith, 2011). When running out of ideas, humans working with the GLM can request GLM ideas. These may act as stimuli, activating the human's memory and triggering the human to themselves develop more own ideas, i.e., cognitive stimulation could occur (Dugosh et al., 2000; Nijstad et al., 2010), increasing the human's fluency (Althuizen & Reichel, 2016; Dugosh et al., 2000) as compared to solitary humans, who do not receive stimuli. However, a human working with a GLM might reduce their effort (i.e., social loafing could occur), particularly if they feel their effort is dispensable or feel less responsible for the results (Pinsonneault et al., 1999). Given that GLMs can outperform many humans in generating ideas (Haase & Hanel, 2023), humans might feel dispensable and might, therefore, reduce their effort. Additionally, previous research suggests that humans might attribute some of the responsibility to the AI (Memmert, 2024; Memmert & Tavanapour, 2023; Stieglitz et al., 2022), potentially leading to a reduction in effort and fewer ideas contributed. While the above suggests both performance-enhancing and -reducing effects, recent studies of multi-human groups suggest that the flow of GLM ideas may make it easier to come up with more ideas (Nomura et al., 2024; Specker et al., 2025; Zhang et al., 2025) and may reduce loafing (Nomura et al., 2024); hence, we propose:

H1-a: Humans brainstorming with a GLM achieve higher fluency (excluding GLM ideas) compared to humans brainstorming alone (solitary brainstormers).

For flexibility or breadth of exploration, different effects could influence performance. Cognitive inertia refers to producing more ideas in the same category (i.e., higher within-category fluency) but exploring fewer categories (i.e., lower flexibility). It can occur when brainstorming alone (here, *human-only* condition) without seeing others' ideas, thereby receiving external stimuli (Pinsonneault et al., 1999), e.g., from a GLM (here, *human+GLM* condition). However, the effect of GLM ideas depends on the degree of relatedness and the amount. Exposure to many closely related ideas might lead to higher conformity to those idea or idea categories, and, consequently, to reduced flexibility (Kohn & Smith, 2011; Lamm & Trommsdorff, 1973). On the other hand, if the stimuli like GLM ideas are diverse, covering different categories, inertia might be reduced (Briggs & Reinig, 2010), aiding humans to explore new categories, i.e., increasing flexibility. Given recent literature suggesting that AI and GLMs, specifically, can expand the problem and solution space (Bouschery et al., 2023; Nomura et al., 2024; Specker et al., 2025), we suggest:

H1-b: Humans brainstorming with a GLM achieve higher flexibility (excluding GLM ideas) compared to humans brainstorming alone (solitary brainstormers).

Stimuli can enhance idea novelty and value (Althuizen & Reichel, 2016). However, exposure to typical or related ideas is associated with reduced idea novelty and increased idea value, and exposure to more ideas with increased fixation (Kohn & Smith, 2011; Wang & Nickerson, 2019). Prior literature comparing human and GLM ideas shows that GLMs can

produce ideas of comparable or even higher creativity compared to many humans (Haase & Hanel, 2023), which should result in humans also generating more novel ideas (Wang & Nickerson, 2019). Results on the impact of GLMs on individual performance are mixed. Kumar et al. (2025) find that when humans work in a short brainstorming session (2 minutes) and receive a limited number of GLM ideas (7), they do not produce more novel ideas. However, Specker et al. (2025) report that when working with a GLM, participants report perceiving increased idea novelty at an individual level. We propose that when participants have more time and can request new GLM ideas without restriction, this high number of creative GLM ideas should induce cognitive stimulation (Leggett Dugosh & Paulus, 2005) and should allow humans to more broadly and deeply explore the topic, which is associated with more novel and valuable ideas (Althuizen & Reichel, 2016). Therefore, we suggest:

H1-c: Humans brainstorming with a GLM produce more novel ideas (excluding GLM ideas) compared to humans brainstorming alone (solitary brainstormers).

H1-d: Humans brainstorming with a GLM produce more valuable ideas (excluding GLM ideas) compared to humans brainstorming alone (solitary brainstormers).

15.2.3.2 Hypothesis Set 2: Dyad Performance

For hypothesis set 2: dyad performance, we compare the overall performance between the conditions, uncovering if superior performance by the human–GLM dyad compared to the individual — a key ambition for human–AI groups (Dellermann et al., 2019) — is achieved.

GLMs “are programmed to create vast amounts of text” (Haase & Hanel, 2023, p. 10) and were shown to be capable of generating ideas (Dell’Acqua et al., 2023; Haase & Hanel, 2023). Humans working with the GLM could repeatedly request and copy GLM ideas, steadily increasing group fluency directly. Thus, human–GLM dyads might outperform solitary humans with respect to fluency. In multi-human groups, GLM-supported groups produced more ideas (Specker et al., 2025). Therefore, we propose:

H2-a: Human–GLM dyads achieve higher fluency compared to solitary humans when brainstorming.

GLMs can produce many ideas, and a higher number of ideas is associated with higher flexibility (Nijstad et al., 2010). However, due to how machine-learning-based AI works, particularly as it is trained on historical data, there is a discussion on whether AI can be creative (Wu et al., 2021) or produces “more of the same”. Indeed, GLMs might reproduce bias and show limited output diversity (Bender et al., 2021; Lin et al., 2022). In a brainstorming context, producing “more of the same” might manifest in the GLM suggesting similar ideas, i.e., taking a narrow perspective on the topic, which aligns with the view of creativity being a comparative human strength (Dellermann et al., 2019). Additionally, research suggests humans working with GLMs to produce ideas with less variability, i.e., potentially lower flexibility (Dell’Acqua et al., 2023). However, recent work on group brainstorming indicates higher flexibility when working with a GLM (Specker et al., 2025). We thus propose:

H2-b: Human–GLM dyads achieve higher flexibility (i.e., breadth of exploration) compared to solitary humans when brainstorming.

Higher fluency and higher flexibility are associated with higher originality (Nijstad et al., 2010). Potentially, with the exploration of less common categories, less common (i.e., more original) ideas are explored (Nijstad et al., 2010). Additionally, a more in-depth exploration of categories is related to higher originality (Althuizen & Reichel, 2016). Thus, the human and the GLM could jointly explore the problem more broadly or more in-depth, both of which is associated with higher originality (Nijstad et al., 2010). Previous research separately investigating humans and GLMs suggests that GLMs can produce ideas of comparable novelty and value to humans or even outperform many humans, though the most novel ideas still originate from humans (Boussieux et al., 2023; Haase & Hanel, 2023). Thus, it was suggested that the human and AI could collectively outperform either entity alone, with humans focusing on producing novel and GLMs on producing valuable ideas (Boussieux et al., 2023). Furthermore, in a human–GLM brainstorming setting, many participants selected GLM ideas as their best ideas (Mommert & Tavanapour, 2023). We thus propose:

H2-c: Human–GLM dyads produce more novel ideas compared to solitary humans when brainstorming.

H2-d: Human–GLM dyads produce more valuable ideas compared to solitary humans when brainstorming.

15.2.3.3 Hypothesis 3: Cognitive Load

For hypothesis 3: cognitive load, we consider how the cognitive load differs across the conditions. Social loafing describes the reduction of effort at the expense of others. Non-human group members may relieve cognitive burdens on humans and support them during task execution (Boussieux et al., 2023; Stieglitz et al., 2022). Therefore, the reliance of the human on AI might be desirable to maintain cognitive resources and improve work efficiency (“smart loafing”, Stieglitz et al., 2022), and previous research shows that humans might save cognitive resources when working with AI (Brachten et al., 2020). It is unclear whether this effect transfers onto divergent thinking tasks like brainstorming, i.e., whether humans brainstorming with GLMs could maintain cognitive resources and would, therefore, report lower levels of cognitive load upon task completion. However, having to pay attention to others’ ideas during brainstorming could result in a distraction effect (Pinsonneault et al., 1999) and potentially increase cognitive load (Coskun et al., 2000) due to the need to process the stimuli and additional knowledge activation (Santanem et al., 2004). Particularly, if the stimuli are of low relatedness, this could increase cognitive load if distant or even topic-irrelevant knowledge (Wang & Nickerson, 2019) is activated. Moreover, prior research suggests that additional cognitive demand is placed on humans when working with GLMs, because they need to evaluate the outputs (Tankelevitch et al., 2024). Research for multi-human brainstorming groups showed that too many GLM hints can increase perceived cognitive load (Zhang et al., 2025) and that participants do not simply

accept GLM ideas, but review and evaluate them (Gonzalez et al., 2024), adding to the cognitive load. We, therefore, propose:

H3: Humans brainstorming with a GLM report a higher cognitive load compared to solitary brainstormers.

15.3 Method

15.3.1 Experimental Design and Procedure

We implemented a single-factor between-subjects experiment with two conditions: human-only vs. human+GLM. The study had four steps: briefing, 10-minute brainstorming (cf. Michinov 2012), selection of top ideas, and post-survey. The procedure is summarized in Figure 4. Task instructions followed Osborn's (1953) rules: we asked participants to come up with as many ideas as possible and encouraged wild ideas. The brainstorming question was: "How can we reduce food waste?", which was used in prior studies (Zhu et al., 2021). Participants in the human+GLM condition could request GLM ideas during the session.

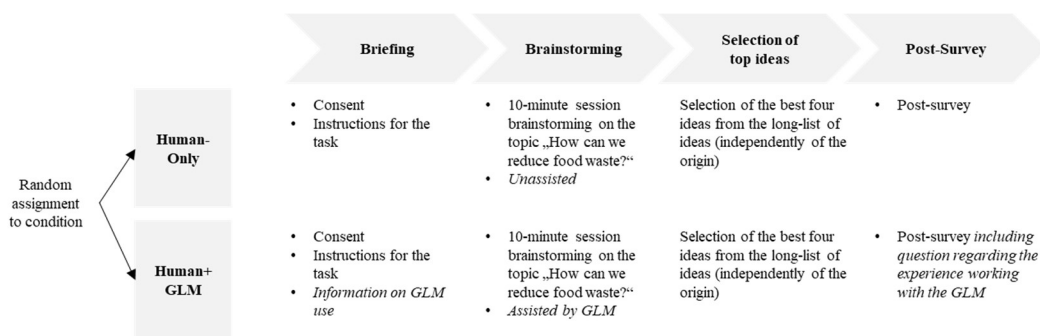


Figure 4 Experimental procedure with differences between conditions marked in italic font

15.3.2 Experimental Treatments

Participants were randomly assigned to one of two conditions. In the *human-only* baseline, each person brainstormed individually with the app's GLM function deactivated. In the *human+GLM* treatment, participants can request and copy GLM suggestions during their brainstorming session whenever they desire; instructions required at least one such request.

There are many alternatives for selecting a baseline, e.g., the human brainstorms alone (ours, comparable to nominal brainstorming groups), with other support systems, or with other humans. We chose the human-only baseline for its research and practical relevance. From the perspective of human-AI collaboration research, the comparison is relevant to explore to what extent the human-AI group achieves "superior" or "complementary" performance (Dellermann et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021) as compared to the individual, a key aspect of hybrid intelligence (Dellermann et al., 2019; Hemmer et al., 2021), which we will seek to quantify. Comparing the performance of individuals with an ICSS to unaided individuals has been typical for brainstorming research to understand

the effects of such tools (Althuizen & Reichel, 2016; Dell’Acqua et al., 2023; Siemon et al., 2015; Wang & Nickerson, 2017). We did not use human groups as the baseline either, as we seek to understand the effect of working with the GLM on cognitive load in alignment with the “smart loafing” construct (Stieglitz et al., 2022). From a practical perspective, other humans for brainstorming might not always be (cost-efficiently) available.

15.3.3 Material: GPT-3.5-based Brainstorming Application and Use Process

To operationalize our treatment and investigate the effects of using GLMs for brainstorming, we developed a brainstorming app without social cues (Figure 5), allowing adding, editing, and removing ideas based on prior research (Di Fede et al., 2022; Memmert & Tavanapour, 2023), following the approach of adapting — through task-specific prompting — and integrating an existing GLM (see Schneider et al., 2024). We asked participants to enter ideas one at a time. Only for participants in the human+GLM condition, the GLM functionality on the right was shown. The list is empty initially. With every click on the “generate ideas” button, three ideas are generated and shown at the top of the list. The copy button allows for accepting the GLM ideas into the long list of ideas (left).

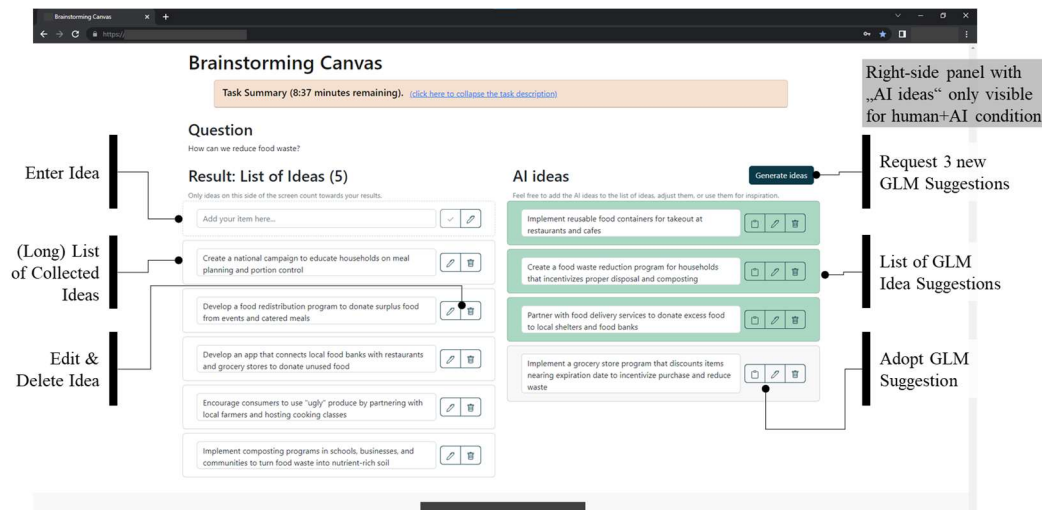


Figure 5 Brainstorming app in the treatment condition (human+GLM) with participant’s long list of ideas (left) and on-demand-generated GLM ideas (right)

Participants had to request GLM ideas (pull) instead of providing them automatically (push), as this was shown to be most effective (Siangliulue, Chan, et al., 2015). When the user clicked the “generate ideas” button, our app’s backend automatically assembled a prompt based on the current canvas state and sent a request to a GLM. The response was processed by our app’s backend and displayed as separate GLM ideas without any changes, e.g., without filtering or randomization. The GLM capability is embedded into the app, abstracting away the interaction with the GLM. This allows humans to focus on the task, freeing them from interacting with GLMs, which might not be easy for non-technical or novice users (Zamfirescu-Pereira et al., 2023). Appendix A (available online via

<http://link.springer.com>) depicts the user's interaction with the app and the information flow to and from the GLM.

As a GLM, we used GPT-3.5-turbo-0301 with a temperature of 0.9 as OpenAI, 2023 recommends for creative tasks. At every "Generate ideas" click, the backend filled a pre-defined prompt template (cf. Memmert, Cvetkovic, & Bittner, 2024; see Figure 6; full explanation in Online Appendix B) with the brainstorming question and the number of ideas per request, and the live canvas state:

- the participant's own idea list (left panel);
- any ideas the GLM had already produced for that participant in the same session (right panel).

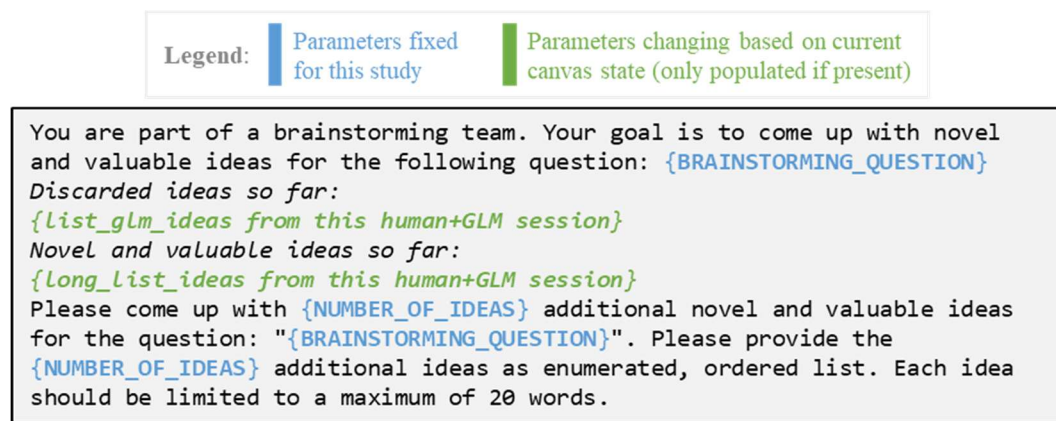


Figure 6 Prompt template with placeholders for producing GLM ideas

Including these lists follows Osborn's (1953) idea of building on others' ideas and might lead the GLM to avoid repeats. No ideas were shared across participants, so sessions remained independent. For each request, the GLM returned three ideas, delivered as an ordered list and limited to 20 words each. The word limit was introduced because pre-tests showed responses to be lengthy otherwise. To ensure a consistent participant experience, if the request to the GLM timed out, we displayed suggestions generated with the initial prompt.

15.3.4 Participants

77 participants brainstormed and completed the survey. We excluded data from two participants: one who misunderstood the task, and one who did not request any GLM ideas despite being in the human+GLM condition. Of the remaining 75 participants (age: mean=23.3, min=18, max=37 years; gender: 13 female, 62 male), all but one are in study programs at the informatics department. The experiment was integrated into in-person university seminars across four courses.

15.3.5 Measures

We measured brainstorming performance via idea fluency and different idea quality dimensions as common brainstorming performance measures (Paulus, 2000). Idea fluency

refers to the number of ideas produced (Paulus, 2000). However, fluency is not meaningful for directly comparing humans and GLMs, because GLMs can autonomously generate large quantities of text with minimal effort (Haase & Hanel, 2023). In our design, by contrast, GLM output is human-bounded: a limited number of ideas are generated only when participants explicitly request them, and these ideas can be added to the idea list only through an active copy action. Consequently, the GLM cannot autonomously inflate fluency through continuous generation. Instead, the number of GLM ideas in a session depends on the participant's choice, making the dyad's output the relevant unit of analysis, as it reflects interaction behavior rather than raw machine capacity.

For quality, we followed the approach of Siangliulue et al. (2015) in having participants select their best ideas after the session, i.e., reducing their long list to a shortlist of four ideas. Shortlisted ideas were evaluated independently by five blind-to-condition judges, including the three authors not involved in app development, on a 7-point Likert scale. We adapted the dimensions and their definitions from Siangliulue (2015), evaluating the ideas according to *novelty* — “consider how novel, original or surprising the idea is” — and *value* — “consider how useful the idea is”. The judges were provided with examples of ideas rated as very/somewhat/not at all novel and valuable, respectively, extracted from a prior study on the same topic (Zhu et al., 2021). We calculated the averages for novelty and value for the shortlisted ideas per participant. Including only the shortlisted ideas ensures that a high number of ideas, which might include low-quality ideas, is not penalized.

GLMs can be used as judges for open-ended questions, an approach known as LLM-as-a-judge (Zheng et al., 2023), and, more specifically, to rate ideas (Dell'Acqua et al., 2023; Haase & Hanel, 2023; Organisciak et al., 2023), showing good correlation with human ratings. Hence, we used GPT-4 in addition to human judges. Adapting an existing approach (Organisciak et al., 2023), we prompted the GLM to rate each idea on novelty and value. We provided the GLM with the same data as the human raters to achieve comparability, including the same definition of novelty and value, the same and exemplary rated ideas, which was shown to increase performance (Organisciak et al., 2023). Results based on GLM ratings are reported separately. Besides considering the participant's self-selected idea shortlist, we used the “Good-Idea-Count” metric, counting ideas exceeding a set rating threshold (Reinig et al., 2007) in the idea long list, enabling an analysis independently of the subjective selection of the participants. As a threshold, we used the average GLM rating for novelty and value, respectively.

Beyond individual ideas, we consider how broadly the topic was covered in each session, so-called flexibility, by categorizing each idea (Althuisen & Reichel, 2016; Nijstad et al., 2010). After reviewing multiple category systems on the brainstorming topic “food waste”, we selected Specht et al.'s system (2019), as its development was based on (1) user-generated content (tweets) of similar length compared to the contributions in our study and based on (2) contributions from a Western country (US; our participants attend a Western Euro-

pean University). Additionally, (3) the system had an appropriate level of granularity (Althuizen & Reichel, 2016). A blind-to-condition student assistant categorized all except for three ideas. Non-categorized and deleted ideas were excluded from the analysis. For each brainstorming session, flexibility was determined as the number of categories with at least one idea present.

We considered both individual- and dyad-level performance (see Figure 2). For the human-only condition, dyad performance is set to individual performance. On the individual level, for the human+GLM condition, only ideas originating from the human were considered. GLM ideas that were not included in the long list by the human were not considered, which was clearly explained in the task introduction.

For measuring perceived cognitive load, we used the same instrument as Stieglitz et al. (2022), the self-report-based, six-dimensional NASA-TLX (Hart & Staveland, 1988). The post-survey contained open- and closed-ended questions (Table 1, Online Appendix C for full questionnaire). The operationalization for all brainstorming performance metrics and variables is detailed in Online Appendix D.

Table 1 Open-ended questions for the human+GLM condition

Felt responsibility (after asking participants to quantify their felt responsibility): Please briefly explain your reasoning for the previous answer for the share of responsibility regarding the result type 'list of ideas' (prior to final idea selection) Please briefly explain your reasoning for the previous answer for the share of responsibility regarding the result type 'list of 4 best ideas' (after the final idea selection)
GLM effect: How did the task change for you due to the presence of the AI (as compared to brainstorming alone)? How did the AI (ideas/suggestions) help you in this brainstorming task? How did the AI (ideas/suggestions) hinder you in this brainstorming task (i.e., made it difficult to perform the brainstorming task)?

15.3.6 Data Collection, Preparation, and Analysis

We use two sources of data: the participants’ interaction data from the brainstorming sessions, such as the number of ideas or the number of suggestion requests, and the survey data, allowing us to contrast actual behavior with the participants’ perception. For the evaluation of the shortlisted ideas, the ideas were randomized, exported, and evaluated. Using the consistency definition with average measures (McGraw & Wong, 1996), the inter-rater reliability ICC(C,k) was .864 for novelty and .736 for value, indicating “strong agreement” (LeBreton & Senter, 2008). Additionally, we adapted an approach of using GPT-4 to rate the ideas (Haase & Hanel, 2023) on the same dimensions and scale. We find a significant positive correlation with high correlation coefficients for both novelty (Spearman $\rho=.726$, $p<.001$) and value ($\rho=.727$, $p<.001$) between average human and GPT-4 ratings, indicating humans and GPT-4 generally agree on idea ratings, which is consistent with prior literature. For statistical analysis, we used the JASP statistics software (JASP, 2023). A sensitivity analysis in G*Power (Faul et al., 2009) using two-tailed tests with $\alpha=.05$, 80 % power and

group sizes of 38 and 37, showed the experiment was sensitive to effects of Cohen's $d \geq 0.656$ for independent-samples *t* tests and $d \geq 0.671$ for Mann–Whitney *U* tests.

To analyze the responses to the qualitative open-ended questions (Table 1), we followed a qualitative content analysis approach (Hsieh & Shannon, 2005; Mayring, 2014), more specifically, a directed content analysis (Hsieh & Shannon, 2005). The goal of a directed approach is to “extend conceptually a theoretical framework,” particularly when “existing theory or prior research exists about a phenomenon that is incomplete or would benefit from further description” (Hsieh & Shannon, 2005, p. 1281). This fits well with our study: while the mechanisms we investigate — cognitive stimulation/inertia, effort (in-)dispensability, responsibility (non-)diffusion, and cognitive load — are well established in human groups, their manifestation in human–AI settings is uncertain. Because these same mechanisms are used to motivate and derive our quantitative hypotheses (see Section 2.3), we used these as deductive categories for our qualitative coding. This alignment allows the qualitative analysis to extend and contextualize the interpretation of the quantitative findings.

After coding approximately 20 % of the responses, two authors reviewed and discussed the coded material to resolve discrepancies and clarify category application. The finalized coding scheme (Online Appendix E) was used to analyze the remaining material. Individual answers to the open-ended survey questions were used as the unit of analysis, and illustrative quotes are reported in the results section.

15.4 Results

Participants contributed 373 ideas, made 206 requests for suggestions, copied 177 suggestions, and selected 283 ideas as their best ideas. Of those ideas selected as best, 224 (79.2 %) originated from humans, and 59 (20.8 %) originated from the GLM. We compared the two conditions on brainstorming performance on two levels (human, dyad) and four measures (fluency, flexibility, novelty, value), and on cognitive load.

15.4.1 Hypothesis Set 1: Human-Level Performance

To measure the GLM's effect on humans, we compared the humans' performance, including only human ideas, but excluding GLM ideas. The normality assumption was not fulfilled for the human-only condition (Shapiro-Wilk: $W=.877$, $p<.001$) for fluency on the human level (human_{fluency}). Thus, we calculated the non-parametric Mann-Whitney-*U* test, which showed a significant difference ($U=475.0$, $p=.015$; Figure 7-a) between the human-only ($mdn=5$) and human+GLM ($mdn=4$) condition. The rank-biserial correlation $r_B=-.324$ suggests this to be a medium effect size (Goss-Sampson, 2022). This corresponds to a Cohen's $d=0.581$ — below the study's sensitivity threshold of $d \geq 0.671$ for Mann–Whitney tests — so the significant difference should be interpreted with caution.

For human-level flexibility, we counted the distinct categories covered per session. The normality assumption was not fulfilled for the human-only condition ($W=.914$, $p=.006$). The Mann-Whitney U test did not show a significant difference ($U=656.0$, $p=.617$; Figure 7-b) for flexibility ($\text{human}_{\text{flexibility}}$).

We compared the shortlisted ideas' qualities on a human level. While for the human-only condition, necessarily all ideas originated from the human, for the human+GLM condition, four participants who did not select any of their own ideas into the shortlist were excluded. Regarding $\text{human}_{\text{novelty}}$, the assumption of normality was not fulfilled for the human+GLM condition ($W=.923$, $p=.022$). The calculated Mann-Whitney-U test was not significant ($U=699.5$, $p=.406$; Figure 7-c). For $\text{value}_{\text{human}}$, the assumption for normality was not fulfilled for the human+GLM condition ($W=.928$, $p=.032$). The test did reveal no statistically significant difference between the groups ($U=737.5$, $p=.204$; Figure 7-d).

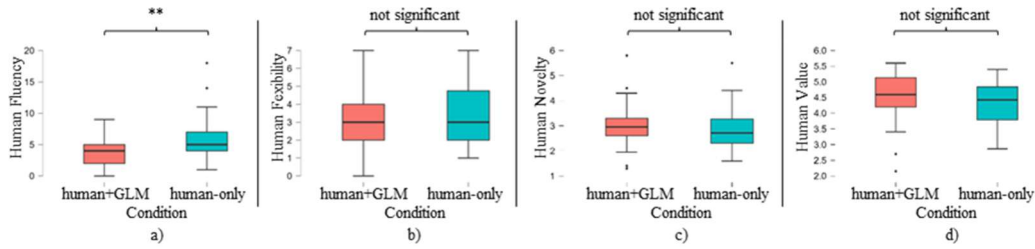


Figure 7 Comparison of human-level performance on
a) $\text{Human}_{\text{fluency}}$, b) $\text{Human}_{\text{flexibility}}$, c) $\text{Human}_{\text{novelty}}$, d) $\text{Human}_{\text{value}}$ ($p < .01$)**

In line with recent studies (Haase & Hanel, 2023; Organisciak et al., 2023), we used GPT-4 (additionally to humans) to rate the ideas and performed the same comparative group analyses. Again, we find no significant differences for novelty ($U=757.0$, $p=.569$) and value ($U=653.5$, $p=.602$) for the shortlisted ideas, underlining the robustness of our findings. Additionally, using the “Good-Idea-Count” metric (Reinig et al., 2007) with GPT-4 ratings, we counted the number of good ideas, i.e., above average ideas for novelty (mean=3.396) and value (mean=4.747), respectively, in the long-lists of ideas. We find no significant differences for novelty ($U=663.5$, $p=.672$) and value ($U=619.5$, $p=.369$), confirming the results.

15.4.2 Hypothesis Set 2: Dyad-Level Performance

For dyad-level fluency ($\text{dyad}_{\text{fluency}}$), the assumption of normality was not fulfilled for both groups (human-only: $W=.877$, $p<.001$, human+GLM: $W=.818$, $p<.001$). The Mann-Whitney U test shows a significant difference between the human-only (mdn=5) and human+GLM (mdn=7) condition on dyad-level *fluency* ($U=963.5$, $p=.006$; Figure 8-a). The rank-biserial correlation $r_B=.371$ suggests this to be a medium to large effect size (Goss-Sampson, 2022).

We then analyzed the flexibility, i.e., the number of distinct categories covered on the dyad level. The assumption of normality was not fulfilled for either of the conditions (human-only: $W=.914$, $p=.006$; human+GLM: $W=.941$, $p=.048$). The Mann-Whitney U test showed

a significant difference ($U=1080.0$, $p<.001$; Figure 8-b) with a large effect size ($r_B=.536$) for $dyad_{flexibility}$ between the human-only ($mdn=3$) and human+GLM ($mdn=5$) conditions.

Lastly, we analyzed the shortlisted ideas according to novelty and value on a group level. For $dyad_{novelty}$, the assumptions of normality and homogeneity of variances (homoscedasticity) were fulfilled. The Student's t-test showed a significant difference with a medium to large effect size ($t(73)=2.889$, $p=.005$, Cohen's $d=.667$; Figure 8-c) for $dyad_{novelty}$ between the human-only ($mean=2.876$) and human+GLM ($mean=3.396$) condition. For $dyad_{value}$, the assumptions of normality were not fulfilled for the human+GLM condition ($W=.797$, $p<.001$). The Mann-Whitney U test revealed a significant difference with a medium to large effect size ($U=994.0$, $p=.002$, $r_B=.414$; Figure 8-d) between the human-only ($mdn=4.425$) and human+GLM condition ($mdn=4.900$) for group level value. The results are shown in Figure 8 with interval plots for the variables following the normal distribution and boxplots for non-normally distributed variables.

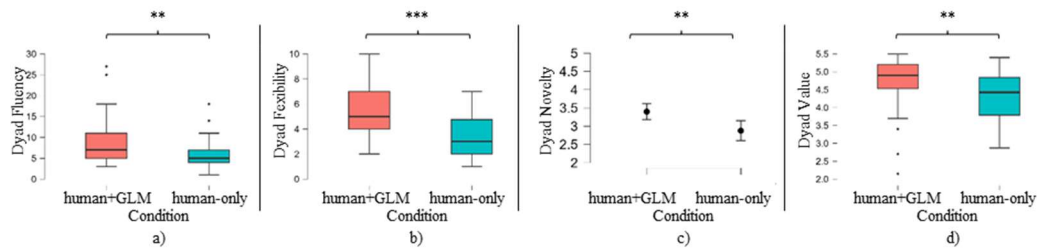


Figure 8 Comparison of dyad-level performance on

a) Dyad_{fluency}, b) Dyad_{flexibility}, c) Dyad_{novelty}, d) Dyad_{value} ($p < .01$, *** $p < .001$)**

Using GPT-4 ratings, we found results similar to the human ratings for the shortlisted ideas, i.e., significant differences between the human-only and human+GLM condition for both novelty (group_{novelty} GPT-rated: $U=932.5$, $p=.015$) and value (group_{value} GPT-rated: $U=986.5$, $p=.003$). Using the “Good-Idea-Count” (Reinig et al., 2007), we find significant differences between the conditions for both novelty ($U=1073.5$, $p<.001$; $mdn_{human-only}=2$, $mdn_{human+GLM}=4$) and value ($U=1095.5$, $p<.001$; $mdn_{human-only}=2$, $mdn_{human+GLM}=6$), increasing result robustness.

15.4.3 Hypothesis 3: Cognitive Load

To understand if humans could reduce their cognitive load by relying on the GLM (cf. Stieglitz et al., 2022), we calculated the Student's T-test. We find no significant difference between the two conditions overall ($t(73)=-.892$, $p=.375$; Figure 9) and for none of the six dimensions (mental demand, physical demand, temporal demand, performance, effort, frustration) of the cognitive load instrument (Cronbach's $\alpha=.762$, Hart & Staveland, 1988).

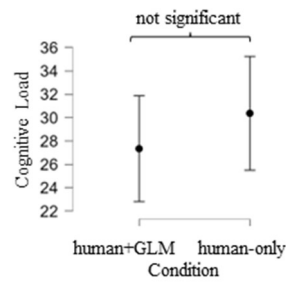


Figure 9 Comparison of cognitive load

In summary, brainstorming performance was significantly lower regarding fluency for humans working with GLMs. We did not observe significant differences for flexibility, novelty, or value for ideas of human origin between conditions. However, we found brainstorming performance to be significantly higher for human–GLM dyads as compared to unaided humans for fluency, flexibility, novelty, and value, i.e., superior group performance (Dellermann et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021) compared to individuals occurred. We find no significant difference between the conditions for cognitive load overall or on individual dimensions. Table 2 summarizes the findings.

Table 2 Findings for human- and dyad-level performance and cognitive load between the conditions human-only and human+GLM

Hypotheses	Level	Variable	Significant difference at 95%-confidence	Condition with higher performance (or, for H3, with higher cognitive load)	
			Actual	Hypothesized	Actual
1-a	Human	Fluency (No. of ideas)	Yes	human+GLM	Human-only
1-b	Human	Flexibility (No. of categories)	No	human+GLM	---
1-c	Human	Novelty	No	human+GLM	---
1-d	Human	Value	No	human+GLM	---
2-a	Dyad	Fluency (No. of ideas)	Yes	human+GLM	Human+GLM
2-b	Dyad	Flexibility (No. of categories)	Yes	human+GLM	Human+GLM
2-c	Dyad	Novelty	Yes	human+GLM	Human+GLM
2-d	Dyad	Value	Yes	human+GLM	Human+GLM
3	Human	Cognitive Load	No	human+GLM	---

15.4.4 Explorative Analysis of Performance-affecting Aspects

To contextualize our findings, we explore performance in the human+GLM condition, particularly regarding the mentioned group effects. First, we explore how the shortlisted ideas differ depending on their origin. Of the 283 shortlisted ideas, 224 were of human and 59 of GLM origin. The normality assumption was not fulfilled for the set of human ideas for both novelty ($W=.945$, $p<.001$) and value ($W=.929$, $p<.001$). The Mann-Whitney-U test showed a significant difference ($U=9654.0$, $p<.001$) between the ideas originating from humans ($mdn=2.8$) and from GLM ($mdn=4.0$) on novelty with a medium to large effect size ($r_B=.461$), and for value ($U=10573.0$, $p<.001$, human: $mdn=4.6$, GLM: $mdn=5.2$) with a large effect size ($r_B=.600$). We found the same significant differences when using GPT-4

ratings. Given that GLM ideas were rated significantly higher, we thus assume a higher share of GLM ideas in the participants' shortlist to be correlated with dyad-level novelty and value ratings. We find a strong positive correlation between the share of GLM ideas on the shortlist and both $\text{dyad}_{\text{novelty}}$ (Person's $r=.458$, $p=.004$) and $\text{dyad}_{\text{value}}$ (Spearman's $\rho=.649$, $p<.001$).

15.4.4.1 Cognitive Stimulation and Cognitive Inertia

We analyzed the responses to the open-ended survey questions to shine a light on potential cognitive stimulation or inertia when working with a GLM and receiving its dynamically generated ideas. In line with prior work (Houde et al., 2025; Memmert & Tavanapour, 2023), several participants stated to have been inspired by the GLM ideas to start generating ideas, to generate additional ideas, or to get unstuck (sic!):

- “It was an inspirational help” (P68) and “It inspire me” (P55)
- “The new ideas of the AI helped me thinkig of additional ideas” (P66)
- “The ideas on the side which came from the AI were a great inspiration for developing new ideas on my own” (P67)

Cognitive stimulation might manifest in humans producing more, or more novel or valuable ideas (Althuizen & Reichel, 2016; Dugosh et al., 2000). However, as described above (hypothesis set 1), humans in the human+GLM condition produced fewer ideas themselves ($\text{human}_{\text{fluency}}$) and did not show higher levels of flexibility, novelty, or value ($\text{human}_{\text{flexibility}}$, $\text{human}_{\text{novelty}}$, $\text{human}_{\text{value}}$), which seems contrary to these subjective reports of feeling inspired. We asked participants about their perceived cognitive stimulation, adapting a 6-item instrument by Gozzo et al., 2022 ($\alpha=.723$). We did not find correlations between self-reported cognitive stimulation and $\text{human}_{\text{fluency}}$, $\text{human}_{\text{flexibility}}$, $\text{human}_{\text{novelty}}$, or $\text{human}_{\text{value}}$ to be significant. Thus, potentially, participants did not actually experience cognitive stimulation but just felt to more efficiently and holistically cover the topic as a (human–GLM) dyad. Indeed, there were significant correlations of reported cognitive stimulation with $\text{dyad}_{\text{fluency}}$ (Spearman's $\rho=.436$, $p=.007$) and $\text{dyad}_{\text{flexibility}}$ ($r=.402$, $p=.014$). From many of the participants' answers, it was unclear if participants felt that the GLM ideas actually stimulated *them* or if these GLM ideas were “just” used as additional ideas for the long list (sic!):

- “It made it easiert to come up with more solutions that I wouldn't have thought of” (P45)
- “AI can help you to bring more ideas to light” (P74)
- “It was easier to generate many ideas” (P41)

Similarly, responses regarding GLM ideas' flexibility were mixed. Some participants stressed that the GLM helped them to gain a new perspective, which could hint at the AI ideas successfully broadening the breadth of exploration and reducing cognitive inertia:

- “The AI included fields that I didn't think about before” (P66)
- “Because of the AI I first came up with certain ideas” (P74)
- “They showed me other fields that I had not considered before” (P63)

However, as reported above, while we observed increased flexibility on the dyad level (dyad_{flexibility}), we did not observe a significant difference on the individual level (human_{flexibility}). Indeed, several participants perceived GLM ideas as repetitive (e.g., P44, P45, P48), with some stating they felt “governed” or their focus was set in a certain direction, which some felt reduced (“cutting off”, P69) or “took away” (P42) their creativity or concentration:

- “It felt a lot faster, but a bit repetitive also since the AI started generating ideas similar to the previous ones” and “[...] It definitely felt like my thought process was being governed by the AI [...]” (P48)
- “The AI stopped coming up with original ideas after a few were generated and I couldn't concentrate on coming up with my own ideas” (P62)
- “[...] set my focus in the direction of the suggestions” (P59)

This suggests that GLM ideas may risk narrowing ideation or reinforcing existing idea paths (cf. Zhan et al., 2024). Although fixation was not universal, this tension between stimulation and narrowing may help explain why individual performance effects were muted, even when some participants reported subjective feelings of stimulation and results surfaced dyad-level gains.

15.4.4.2 Free Riding and Smart Loafing

Reduced effort and disengagement: Regarding free riding, qualitative responses indicated that some participants reduced their effort or became disengaged, for example, stating:

- „I became lazy“ (P61) and “I didn't do much” (P61)
- “By simply leaving [the AI] as an option, I didn't see the need to make an effort to come up with new ideas” (P46)
- “I feel like using the AI made my brain shut off a little bit” (P47) and “It took the "thinking" out of my hand” (P47)
- “I had to do less. [...]“ (P50)

Dispensability of effort: A contributing factor for free riding is felt dispensability of effort (Pinsonneault et al., 1999), i.e., the feeling that (additional) own effort would not (substantially) improve results. In line with recent literature finding that GLMs can produce creative ideas comparable to many humans (Haase & Hanel, 2023), several participants expressed a sense of inferiority compared to the GLM when it came to generating ideas, e.g., describing themselves as “not that creative” (P60) or stating that GLM ideas were “overall better defined and closer to the actual topic” (P42). However, other participants described the

GLM ideas as of “lower quality and more repetitive” (P44), “uncreative and boring” (P51), “too mundane and common” (P75), or as “generic answers already are established in a way” (P73) that would “create different problems” (P53), and stating that their “ideas are unique and make more sense to [them] than the ‘cheap’ AI suggestions” (P49). These mixed results may hint at underlying individual differences for working with GLMs in creative tasks (cf. Memmert, Mies, & Bittner, 2024).

Diffused responsibility: An additional aspect related to free-riding is diffused responsibility (Latané et al., 1979; Pinsonneault et al., 1999). In the post-survey, we asked participants about the degree of responsibility they felt for the longlist and shortlist of ideas. We find that none of the participants claimed full responsibility for the long list, and only 6 participants (16.2 %) for the shortlist of ideas. We found a strong correlation between the felt responsibility and the share of ideas originating from the human in the long list ($r=.710$, $p<.001$). The same holds true for the share of human ideas on and felt responsibility for the shortlist ($r=.489$, $p=.002$). This is also reflected in the qualitative responses, where many participants provided a rationale based on the contribution share for the level of felt responsibility, e.g., stating “I shared a low responsibility because I choose to include comparatively many ideas from the AI.” (P44).

Effort shift and cognitive load: Tying our findings back to theory, smart loafing refers to “the reduction of effort [...] to maintain cognitive resources and enhance efficiency in work” (Stieglitz et al., 2022). While some participants reported having reduced their effort as described above, other participants offered more nuanced responses (sic!):

- “The AI made the task easier and less mentally straining, I didn't have to put all of my focus on coming up with ideas, only on reading and evaluating the AI's ideas” (P62)
- “It was less Brainstorming, more likely a reading task where i chose the best options” (P60)
- “i had a helping hand which never ran out of ideas, so it was more stress relieving.” (P68)

Rather than an effort reduction, these responses signal an effort shift from generating to curating ideas. This effort shift seemed to have been both a source of feelings of indispensability and non-diffused responsibility according to some participants, e.g., stating “[...] I think I wouldnt really rely on AI ranking” (sic!, P47) for the former, and “I evaluated all of the ideas on my own an[d] picked the what i thought were the best ideas” (P68) for the latter. When asked about how to achieve the best brainstorming performance with our app, most participants ($n=30$) selected a mix between generating and curating ideas (see Figure 10). As participants in the human+GLM condition spent time reviewing and selecting GLM ideas, this might explain their lower fluency. Indeed, the number of suggestion requests and $human_{fluency}$ is negatively correlated ($r=-.419$, $p=.01$).

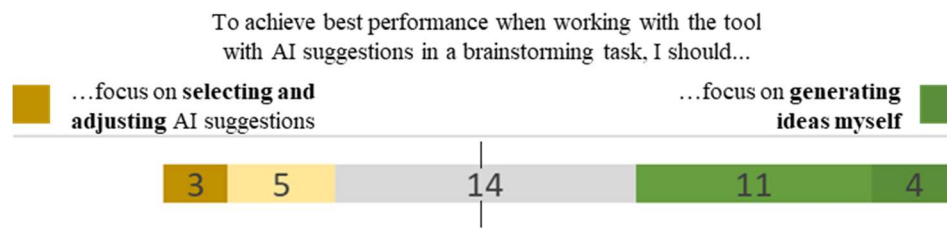


Figure 10 Self-reported optimal effort distribution

This effort shift may explain why we did not find a significant difference in cognitive load, contrary to expectations under the smart-loading assumption. Though several participants reported to have felt working with the AI made the task “easier” (e.g., P41, P42), “less mentally straining” (P62), and “took the stress away” (P59), many reported to having to read and evaluate GLM ideas, explaining “I had to reject the ai suggestions and pay attention not to be influenced too much by it or to be bogged down in the ai's way of thinking” (P75). While not explicitly stated by many as additional cognitive load, this constant evaluation effort might have offset the load reduction due to the GLM contributing ideas (cf. Tankelevitch et al., 2024). Thus, while working with GLMs can offer performance improvements, there are also risks associated.

15.5 Discussion

15.5.1 General Discussion

We explored the effect of a GLM-based ICSS on brainstorming performance and cognitive load through the lens of group mechanisms known from purely human groups. We find that when excluding GLM ideas, *humans* in the human–GLM dyad produce fewer ideas compared to solitary brainstormers (hypothesis set 1). However, *human–GLM dyads* collectively, i.e., including GLM ideas, produce more, more novel, and more valuable ideas and cover more categories, compared to solitary brainstormers (hypothesis set 2). There was no difference in cognitive load between the conditions (hypothesis 3).

15.5.1.1 Superiority of Human–GLM Brainstorming

A key goal for human–AI collaboration is for the human–AI group to achieve superior performance compared to the human; that is, working with the AI system should yield better results than working alone (Dellermann et al., 2019). Similarly, Hemmer et al. (2021, p. 2) and Hemmer et al. (2025) describe the desired outcome for a human and an AI — which they label team members — as working together to achieve “complementary team performance (CTP),” defining it as “the team performance exceed[ing] the maximum performance of both individual entities.” AI was traditionally discussed as being helpful for (automating) repetitive tasks (Krogh, 2018), whereas creative tasks were thought to be in the realm of humans (Dellermann et al., 2019; Redifer et al., 2021), particularly as AI systems’ output was thought to potentially lack diversity (Bender et al., 2021). After the creativity of GLMs and humans was investigated in isolation (Boussieux et al., 2023; Haase & Hanel, 2023; Organisciak et al., 2023), we studied joint human–AI brainstorming, as has

been called for in the literature (Di Fede et al., 2022; Liu et al., 2023; Rafner et al., 2023; Wu et al., 2021). In empirically demonstrating the superiority of the human–AI dyad in brainstorming, i.e., achievement of complementary performance (Hemmer et al., 2025; Hemmer et al., 2021) of the human–AI dyad compared to the individual human, we confirm what was hypothesized in the literature (Boussieux et al., 2023). Our results partially confirm earlier findings, such as those showing heightened fluency and flexibility when brainstorming with AI (Demir et al., 2024; Specker et al., 2025), but contradict other findings, e.g., those indicating reduced novelty (Demir et al., 2024; Kumar et al., 2025). Overall, the results align with the recent meta-analysis by Holzner et al. (2025), which suggests that human–AI dyads outperform solitary humans in creative performance, particularly when participants are non-domain experts and the AI is a capable GLM like GPT-3.5. However, our study offers a more nuanced analysis, distinguishing among four brainstorming metrics rather than using a single aggregated measure of creative performance. Holzner et al. (2025) report that AI reduces the diversity of ideas, whereas our study shows increased flexibility of human–AI dyads as compared to solitary humans. We find that human–AI dyads cover more categories of ideas compared to solitary humans. This may hint at the trade-off that although individuals may benefit from AI usage to increase the human–AI dyad’s performance, on an aggregated level — when comparing pooled ideas of solitary humans and pooled ideas of human–AI dyads — results may become more homogeneous (cp. Dell’Acqua et al., 2023). Future research should explore the differences in measurement approaches, for example, by comparing flexibility through assessing the number of categories covered for a topic (as in our study) with statistical approaches, such as embedding-based measures like cosine similarity (e.g., Dell’Acqua et al., 2023; Holzner et al., 2025).

Moreover, future research should extend performance benchmarking to include both human–human dyads and AI-only conditions to provide a more comprehensive comparison of creative performance across different collaboration types. Including a human–human condition would enable disentangling the effects of GLM support from the general effects of working with another idea-contributing partner. While we deliberately omitted a human–human group to isolate the cognitive and behavioral mechanisms of human–AI interaction, thereby avoiding human social dynamics, future work could systematically contrast these settings. Such comparisons would directly respond to calls for studying human–AI complementarity (Dellermann et al., 2019; Hemmer et al., 2025) and to recent findings that challenge the notion that human input necessarily has a positive performance impact (Lee & Chung, 2024; Vaccaro et al., 2024).

15.5.1.2 Unexpected Effect of GLM-Based ICSS on Human Performance

Intuitively, the results may not come as a surprise — tool support leads to better performance. However, our analysis shows that when excluding GLM ideas, humans working with GLMs themselves do not perform better compared to solitary brainstormers; instead, they perform worse. This is contrary to expectations for traditional stimuli-based ICSS

(Wang & Nickerson, 2017). After all, stimulating the human to create more and more novel ideas is the goal of a stimuli-provider ICSS; i.e., ICSS-induced cognitive stimulation should manifest in higher individual brainstorming performance (Pinsonneault et al., 1999).

Siangliulue, Chan, et al. (2015, p. 88) demonstrated that stimuli-based ICSS are most effective when humans request stimuli themselves (i.e., pull; as in our app), and they showed that “participants primarily requested examples when they ran out of ideas.” While there is no agreed-upon duration for brainstorming, durations in experiments between 4, 10 (this study), and 15 minutes are common (Dennis et al., 2013; Michinov, 2012; Pinsonneault et al., 1999; Siangliulue, Arnold, et al., 2015), and past research has shown that fluency rapidly declines after the first few minutes (Baruah & Paulus, 2016; Kohn & Smith, 2011), with participants adopting more GLM ideas than producing ideas themselves after about 3 to 4 minutes (Memmert, Mies, & Bittner, 2024). Given those empirical timings, participants in our study had sufficient time to both contribute their own ideas and interact with the ICSS when running out of ideas.

In our study, the ICSS provides stimuli in the form of full ideas rather than only cues, the latter of which is common for prior stimuli-provider ICSS. In that, working with our tool might be somewhat similar to brainstorming with another human, who also contributes full ideas (see Figure 1). According to the literature, both incomplete ideas from ICSS and complete ideas from other humans can induce cognitive stimulation and consequently improve human brainstorming performance (Pinsonneault et al., 1999; Wang & Nickerson, 2017). However, contrary to those expectations, here, external stimuli (i.e., GLM ideas) did not increase humans’ brainstorming performance, but rather resulted in the humans generating fewer ideas themselves. Thus, while both the traditional and the GLM-based ICSS resulted in better collective group-level performance, different mechanisms seem to drive these improvements.

What seems interesting is that despite participants objectively not producing more (novel) ideas when working with the GLM, they — in line with prior research (Specker et al., 2025) — subjectively reported feeling stimulated or inspired, which could be rooted in the new possibility that brainstorming with a GLM provides. Unlike with traditional ICSS, which require the human to develop a full idea from the stimuli, with our app providing complete, creative ideas — GLM ideas were rated as more novel than human ideas — these GLM ideas can simply be copied by the human without additional effort. While this is similar to brainstorming with another human, participants did not have to credit another person for ideas they had not developed themselves. Hence, humans potentially felt subjectively inspired and indeed perceived themselves as having covered the topic more broadly because they subconsciously attributed the GLM ideas to themselves. This perspective is compatible with the theory of the “extended self” (Mirbabaie et al., 2021) and with viewing GLM-based ICSS as merely tools.

Perhaps, this reduced the drive of humans to come up with more ideas themselves based on the GLM's ideas, which might mean that the dyad-level performance improvement is mainly driven by the GLM's ideas, rather than by improved human performance. This raises the question of whether “collaborative complementarity potential” (Hemmer et al., 2025, p. 7) is realized — that is, not merely that the dyad outperforms either entity alone, but that their interaction yields performance exceeding the additive combination of their individual contributions — or, put simply, whether the whole exceeds its parts.

15.5.1.3 Smart Loafing Perspective on Performance

Another potential explanation for us not observing higher performance with increased cognitive stimulation could be that the performance-enhancing and -reducing effects canceled each other out. Liu et al. (2023) raised the question, “Will humans be Free-Riders?”. Indeed, we observe humans working with the GLM to contribute fewer ideas (hypothesis set 1), and some admitted to having become “lazy,” suggesting that free riding might have occurred. A cause for free riding is the phenomenon of diffused responsibility. Indeed, no participant felt full responsibility for the long list of ideas, and only a few for the shortlist. Such responsibility attribution to non-human group members has been observed before (Stieglitz et al., 2022). Thus, one might assume that humans reduced their effort, i.e., free rode (Latané et al., 1979). However, on a group level, human+GLM dyads performed better (hypothesis set 2), so task performance was not harmed. We thus concur with Stieglitz et al. (2022) that free riding or social loafing may need to be conceptualized differently in human–AI groups, as it may sometimes be desirable, particularly if it does not lead to a reduction in the quality of the result. Below, we reflect on their concept of smart loafing, i.e., humans' *effort reduction to maintain cognitive resources* while enhancing *work efficiency* (Stieglitz et al., 2022). We discuss the key parts of the definition below.

Though perhaps occasionally, overall, it is not apparent that participants *reduced their effort* when working with the GLM. Humans in the human-only condition focused exclusively on generating ideas, whereas humans in the human+GLM condition split their effort between generating and curating (requesting, evaluating, and adjusting) ideas, implying a difference in roles from pure creator to creator and curator (cf. Gonzalez et al., 2024; Tankelevitch et al., 2024). Perhaps there was not a reduction in effort but a shift in effort allocation.

Maintaining cognitive resources could mean conserving them when performing the task due to the AI. We would then expect lower cognitive load levels for humans working with the GLM. It could also mean improving performance while not increasing cognitive load, i.e., enhancing *work efficiency*. Work efficiency enhancement was included in the original definition (Stieglitz et al., 2022) and was observed in our study, i.e., with similar cognitive load, a higher collective task performance was achieved by the human–GLM dyad. We did not observe lower cognitive load, i.e., humans did not save cognitive resources when working with the GLM (hypothesis 3). A potential explanation is that there was a shift instead

of a reduction in cognitive load. According to cognitive load theory, cognitive load comprises intrinsic, extraneous, and germane components (Paas et al., 2003; Sweller, 1988). While the intrinsic cognitive load for generating own ideas might have decreased, an additional intrinsic cognitive load for curating ideas might have occurred (Figure 11). Confirming earlier reports of such behavior (Gonzalez et al., 2024), many participants reported having reviewed the GLM ideas before adopting them. These findings align with prior research suggesting that GLMs may introduce additional cognitive demands, for example, for the evaluation of AI outputs (Tankelevitch et al., 2024). Moreover, additional extrinsic cognitive load due to the presentation of the GLM ideas may have occurred, and germane cognitive load — the mental effort required for constructing, refining, or integrating knowledge structures or schemas (Paas et al., 2003, p. 2) — may have been maintained. Together, this may explain why we did not observe a reduction in cognitive load.

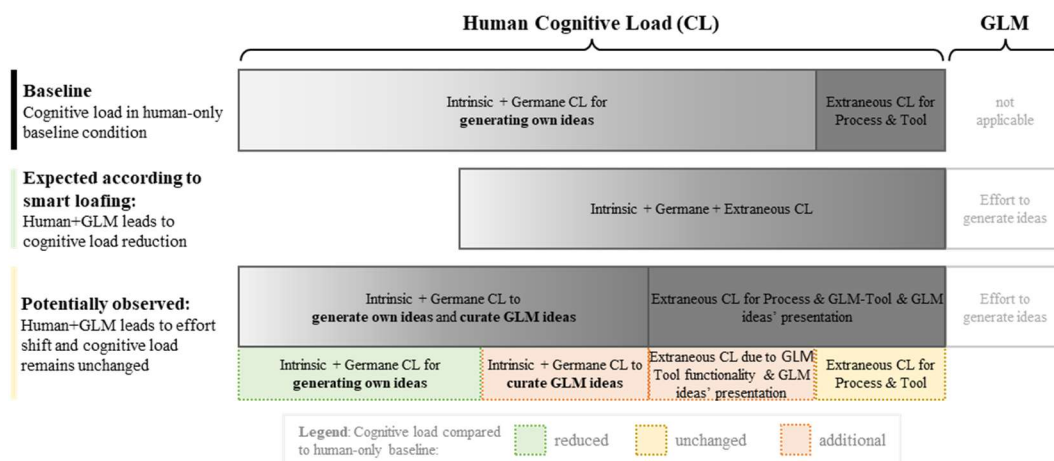


Figure 11 Schematic comparison of the aspired (reduction) and potentially observed (shift) effect on cognitive load when brainstorming with a GLM

Future research might differentiate cognitive load types (Cheng et al., 2020; Javadi et al., 2013) and explore if GLM ideas can be controlled to balance the positive effects of offering ideas and the negative effects of inducing cognitive load, e.g., by controlling idea relatedness (Baruah & Paulus, 2016; Wang & Nickerson, 2017, 2019), e.g., through prompt design or filtering (Summers-Stay et al., 2023) of GLM ideas, or by delivery mechanisms (Siangliulue, Chan, et al., 2015).

Smart loafing implies a deliberate and intelligent approach to loafing. This may require reflection on one's own and the GLM's comparative task-relevant capabilities, such as creative ability or topic knowledge, to ultimately achieve a symbiotic collaboration and a productive division of work (Benbya et al., 2024; Boussioux et al., 2023; Vaccaro et al., 2024). We found signs of this reflection in some qualitative responses. In the literature, various roles have been proposed for AI in human–AI work (Siemon, 2022). Perhaps similarly interesting is the human's role expansion from creator to curator (Memmert et al., 2025). It might be fruitful to structure human roles and support individuals in taking a good role or effort split, i.e., striking a balance between creating and curating in open-ended settings, enabling them to loaf “smartly”.

A related aspect is the attribution of responsibility in human–AI settings. As in prior studies (Stieglitz et al., 2022), participants did not attribute full responsibility to themselves despite being the only human. The share of human ideas was strongly positively correlated with felt responsibility. While causation is unclear, this might hint at humans feeling less responsible the more the AI contributes, even when in control of result curation. With GLMs becoming more capable, this may introduce problems. Future research should explore the rationale for responsibility attribution, such as the perception of the AI on the spectrum from tool to colleague, to inform interventions for calibrating felt responsibility.

Smart loafing was suggested to be desirable for repetitive or overhead tasks (Stieglitz et al., 2022). We show that it might be desirable even for one-off tasks, such as brainstorming for a specific problem, and thus agree with the call for extending the focus of smart loafing beyond its initial, more narrow scope towards other, more cognitively challenging tasks.

15.5.2 Contributions to Literature and Theory

With our study, we contribute to the long history of literature on ICSS and group performance mechanisms (Althuizen & Reichel, 2016; Gabriel et al., 2016; Wang & Nickerson, 2017) as well as to the emerging information systems human–AI collaboration literature, particularly regarding human–AI (superior) performance (Dellermann et al., 2019; Hemmer et al., 2021), group dynamics (Liu et al., 2023; Makarius et al., 2020), and roles (Simon, 2022).

Following up on prior research comparing brainstorming of humans and GLMs in isolation (Haase & Hanel, 2023; Koivisto & Grassini, 2023), and human groups with and without GLM support (Specker et al., 2025), we investigate joint human–GLM brainstorming as suggested, e.g., by Di Fede et al. (2022) and discussed by Holzner et al. (2025). Achieving superior or complementary performance of the human–AI group compared to the individual human is a core ambition for human–AI collaboration (Dellermann et al., 2019; Hemmer et al., 2025; Hemmer et al., 2021) and for stimuli-based ICSS. We empirically demonstrate that such superior or complementary performance can be achieved with the proposed GLM-based ICSS. This is relevant, as we show that such superior performance can be achieved not only in large-scale decision-making (Hemmer et al., 2025; Hemmer et al., 2021), but also in generating ideas, a creative activity traditionally considered a human comparative strength (Dellermann et al., 2019).

Our findings align with recent meta-analytic evidence by Holzner et al. (2025) and seem to confirm prior findings for stimuli-provider ICSS (Wang & Nickerson, 2017). We show that a GLM-based ICSS can enhance brainstorming performance. However, our in-depth analysis reveals a contradiction: humans produced fewer ideas — rather than more, as would typically be expected when external stimuli are present and cognitive stimulation is induced — yet they nonetheless reported feeling subjectively inspired. This suggests that different mechanisms drive performance improvement at the group level in traditional ICSS and GLM-based ICSS, and might hint at undesired side effects. We offer potential explanations

based on the difference in capability and interaction: the GLM contributes complete ideas instead of cues, but the human needs to review those ideas, which may induce additional cognitive load (Tankelevitch et al., 2024). We highlight that mechanisms from human groups, which are increasingly used to investigate human–AI groups (Siemon et al., 2015; Stieglitz et al., 2022), need to be applied carefully to avoid drawing incorrect conclusions. To that end, we offer a nuanced perspective on the “smart loafing” construct (Stieglitz et al., 2022) based on our findings.

15.5.3 Implications for Practice

While free-text, conversational GLM interactions, such as those with ChatGPT, are possible, prompting can be challenging for novices (Zamfirescu-Pereira et al., 2023). Apps can make GLM capabilities available by automatically constructing prompts, abstracting away the GLM interaction, which could be referred to as *embedded GLM capabilities*. However, this might change user behavior in undesired ways, or, as one participant put it, “I became lazy” (P49). Users might overly rely on the suggestions without critically engaging with them (Stieglitz et al., 2022). Such overreliance might be particularly problematic if it results in creative humans, or humans who possess unique knowledge not accessible to the GLM, contributing fewer ideas, particularly because the most creative ideas still originate from humans (Haase & Hanel, 2023). Developers should, therefore, focus on sociotechnical design to address undesired behavioral effects, thereby preventing productivity losses (Alavi et al., 2024) and fostering a symbiotic collaboration between GLMs and humans (Benbya et al., 2024; Feuerriegel et al., 2024). Tool designers may explore, e.g., providing only partial suggestions to “force” humans to actively engage with GLM ideas (Buçinca et al., 2021), which is a prerequisite for cognitive stimulation (Dugosh et al., 2000); change the timing of GLM ideas to encourage humans to think for themselves before working with a GLM, as suggested by some participants, or use dynamic feedback to encourage re-combining and improving ideas (cf. Di Fede et al., 2022; Gordetzki et al., 2023; Javadi et al., 2013), which is a core idea of brainstorming (Osborn, 1953).

Our findings have implications for managing employees who use GLMs. As we observe significant improvements in performance (hypothesis set 2), managers may want to encourage employees to explore the potential of GLMs for their work (Dell’Acqua et al., 2023). However, as we observe a strong correlation between the human contribution share and felt responsibility alongside responsibility attribution to the GLM, managers could educate employees about such expected changes in felt responsibility and consider clarifying their stance on responsibility for work results. Additionally, managers should consider this new way of working when assessing the employees’ performance, for example, by considering both critical engagement with the GLM and the outcome.

15.5.4 Limitations

Our study has several limitations. The sensitivity analysis showed that our sample size was adequate for detecting medium-to-large effects but underpowered for smaller ones; therefore, both the non-significant outcomes and the significant human fluency effect (Cohen's $d=0.581$, marginally below the 80 % power threshold) should be interpreted with caution. We therefore provide qualitative interview data to contextualize both significant and non-significant quantitative findings. In addition, the sample — composed primarily of informatics students — limits generalizability, as such students may have greater prior exposure to and comfort with GLMs, potentially influencing their interaction behavior. Future research should replicate and extend these findings with larger and more diverse samples. Moreover, our participants worked on a general societal question that required no specific knowledge. It might be worthwhile to explore whether and how the participant-task-fit influences GLM usage behavior, for example, by varying the required level of task expertise and the participants' level of expertise, such as assigning programming problems to informatics students. This might be particularly interesting because prior research has highlighted the potential moderating effect of domain expertise (Holzner et al., 2025).

Our study was limited to a specific understanding of brainstorming performance — specifically, idea fluency, flexibility, novelty, and value — but other dimensions exist (Dean et al., 2006). We selected these dimensions because they are commonly used in brainstorming research (Althuizen & Reichel, 2016; Paulus, 2000; Reinig et al., 2007). For group effects, we covered cognitive stimulation, cognitive inertia, and free riding (with smart-loafing adaptation, Stieglitz et al., 2022). While there are more potential effects (Pinsonneault et al., 1999), previous exploratory research has revealed those effects to be relevant for this GLM-based brainstorming setting (Memmert & Tavanapour, 2023). Additionally, we have focused on dyads of one human and one GLM to investigate effects in isolation. However, GLMs will likely impact dynamics when added to multiple-human brainstorming groups (cf. Gonzalez et al., 2024; Muller et al., 2024), which future research should investigate.

Our brainstorming question addresses a societal problem, and such problems in general, and this question in particular, have been used in prior brainstorming research (Zhu et al., 2021). While we did not investigate other questions — limiting generalizability — our technical approach was not tailored to the specific question. On the contrary, the question could easily be replaced in our prompt template. Prior research has shown that GLMs can generate creative ideas across various topics (Haase & Hanel, 2023); nonetheless, future research should investigate the robustness of these results across different topics.

Lastly, our study relied on GPT-3.5 as the embedded GLM, which represented the state of the art at the time of data collection. Generative AI is a rapidly evolving field, and newer models have been shown to outperform earlier ones, including in creative idea generation when assessed in isolation (Haase and Hanel 2023). However, such technical advances do not necessarily translate into improved human–AI creative performance, as shown in the

meta-analysis by Holzner et al. (2025), i.e., increases in model capability may change rather than uniformly enhance co-creative outcomes. Nevertheless, the behavioral mechanisms we identify, such as the redistribution of effort, changes in responsibility attribution, and potential cognitive load trade-offs, may persist across model generations (cf. Memmert & Tavanapour, 2023).

15.5.5 Outlook

We observed humans assuming the roles of both creators and curators. Future research should further investigate group dynamics and roles when working with GLMs. This may include checking whether patterns for GLM usage emerge, how and in which moments mechanisms like cognitive stimulation unfold, and how cognitive load is affected in these situations. To that end, it may be beneficial to combine interaction log data with observational or psychometric measures. With the increased embedding of GLM capabilities in apps, supporting humans in finding the right role balance and adapting favorably (see Benke et al., 2024) will become crucial. Along with taking on multiple roles, we observed that participants did not fully attribute responsibility to themselves. It will be important to understand the effect on behavior and how responsibility attribution can be influenced. Additionally, it will be important to understand how these group mechanisms change when systems with (higher) agency are introduced, e.g., if the human does not have to trigger the GLM but it contributes automatically (cf. Specker et al., 2025).

Regarding how generative AI will affect jobs (Dell’Acqua et al., 2023), we show that not only repetitive tasks (Dellermann et al., 2019) but also creative tasks can benefit from generative AI. However, creativity ranges from everyday creativity to genius-level creativity (Haase & Hanel, 2023), underscoring the need for further research on tasks and people. For the former, the question arises as to which open-ended (creative) tasks can benefit from GLM support. For the latter, the question arises as to who benefits from such tools (Benbya et al., 2024; Dell’Acqua et al., 2023), i.e., do creative and less creative individuals benefit similarly (Wang & Nickerson, 2017)? Lastly, future research should explore the broader implications of GLM usage, both for organizations and for individuals. While we find that individuals benefit in the short term by achieving higher creative performance on a dyad level, recent research suggests that, when taken together, GLM-assisted humans produce less diverse ideas compared to unaided humans (Holzner et al., 2025). Moreover, in the long term, individuals may exhibit a reduced level of engagement and creative thinking ability when using GLMs (Kosmyna et al., 2025; Kumar et al., 2025), raising the question of how tools should be designed for leveraging the creative potential of GLMs while mitigating negative long-term effects on engagement and creative ability.

15.6 Conclusion

As GLMs become more widely adopted, new opportunities for human–AI collaboration in solving complex problems emerge. To ensure the effectiveness of such collaboration, it is

crucial to understand the group dynamics when working alongside AI (Benbya et al., 2024; Makarius et al., 2020). In our experiment, we observed higher performance by the human–GLM dyads compared to the solitary humans. However, humans working with the GLM contributed fewer ideas themselves, perhaps because they had to balance the roles of creator and curator of AI ideas. This role split was accompanied by most participants attributing less than full responsibility to themselves, despite being the only human in the group. We believe future research should explore this role change and support striking a balance between roles when individual human work is transformed into collaborative human–AI work.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s12599-025-00974-y>.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

Author Contributions

Contributions are listed according to the Contributor Role Taxonomy (CRediT, NISO CRediT Working Group): Lucas Memmert: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review & editing. Izabel Cvetkovic: Data curation. Navid Tavanapour: Data curation, Investigation. Eva Bittner: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

15.7 References

- Alavi, M., Leidner, D. E., & Mousavi, R. (2024). Knowledge Management Perspective of Generative Artificial Intelligence. *Journal of the Association for Information Systems*, 25(1), 1–12. <https://doi.org/10.17705/1jais.00859>
- Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of Management Information Systems*, 33(1), 11–44. <https://doi.org/10.1080/07421222.2016.1172439>
- Baruah, J., & Paulus, P. B. (2016). The Role of Time and Category Relatedness in Electronic Brainstorming. *Small Group Research*, 47(3), 333–342. <https://doi.org/10.1177/1046496416642296>
- Benbya, H., Strich, F., & Tamm, T. (2024). Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work. *Journal of the Association for Information Systems*, 25(1), 23–36. <https://doi.org/10.17705/1jais.00861>

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM. <https://doi.org/10.1145/3442188.3445922>
- Benke, I., Knierim, M., Adam, M., Beigl, M., Dorner, V., Ebner-Priemer, U., Herrmann, M., Klarmann, M., Maedche, A., Nafziger, J., Nieken, P., Pfeiffer, J., Puppe, C., Putze, F., Scheibehenne, B., Schultz, T., & Weinhardt, C. (2024). Hybrid Adaptive Systems. *Business & Information Systems Engineering*, 66(2), 233–247. <https://doi.org/10.1007/s12599-024-00861-y>
- Bouschery, S. G., Blazevic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2), 139–153. <https://doi.org/10.1111/jpim.12656>
- Boussiou, L., N. Lane, J., Zhang, M., Jacimovic, V., & Lakhani, K. R. (2023). The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing. *Harvard Business School*. Advance online publication. <https://doi.org/10.2139/ssrn.4533642>
- Brachten, F., Brünker, F., Frick, N. R. J., Ross, B., & Stieglitz, S. (2020). On the ability of virtual agents to decrease cognitive load: an experimental study. *Information Systems and E-Business Management*, 18(2), 187–207. <https://doi.org/10.1007/s10257-020-00471-7>
- Briggs, R. O., & Reinig, B. A. (2010). Bounded Ideation Theory. *Journal of Management Information Systems*, 27(1), 123–144. <https://doi.org/10.2753/MIS0742-1222270106>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Cheng, X., Fu, S., Vreede, T. de, Vreede, G.-J. de, Seeber, I., Maier, R., & Weber, B. (2020). Idea Convergence Quality in Open Innovation Crowdsourcing: A Cognitive Load Perspective. *Journal of Management Information Systems*, 37(2), 349–376. <https://doi.org/10.1080/07421222.2020.1759344>
- Coskun, H., Paulus, P. B., Brown, V., & Sherwood, J. J. (2000). Cognitive stimulation and problem presentation in idea-generating groups. *Group Dynamics: Theory, Research, and Practice*, 4(4), 307–329. <https://doi.org/10.1037/1089-2699.4.4.307>

- Dean, D., Hender, J., Rodgers, T., & Santanen, E. (2006). Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, 7(10), 646–699. <https://doi.org/10.17705/1jais.00106>
- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Demir, S., Fuegener, A., Gupta, A., & Weinmann, M. (2024). The effect of AI support on Torrance’s creativity dimensions: Evidence from an online experiment. In *ECIS 2024 Proceedings*. https://aisel.aisnet.org/ecis2024/track09_coghbis/track09_coghbis/17/
- Dennis, A. R., Minas, R. K., & Bhagwatwar, A. P. (2013). Sparking Creativity: Improving Electronic Brainstorming with Individual Cognitive Priming. *Journal of Management Information Systems*, 29(4), 195–216. <https://doi.org/10.2753/MIS0742-1222290407>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Pers. Soc. Psychol.*, 53(3), 497–509. <https://doi.org/10.1037/0022-3514.53.3.497>
- Dugosh, K. L., Paulus, P. B. [P. B.], Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *J. Pers. Soc. Psychol.*, 79(5), 722–735. <https://doi.org/10.1037/0022-3514.79.5.722>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Gabriel, A., Monticolo, D., Camargo, M., & Bourgault, M. (2016). Creativity support systems: A systematic mapping study. *Thinking Skills and Creativity*, 21, 109–122. <https://doi.org/10.1016/j.tsc.2016.05.009>
- Gonzalez, G. E., Moran, D. A. S., Houde, S., He, J., Ross, S. I., Muller, M. J., Kunde, S., & Weisz, J. D. (2024). Collaborative Canvas: A Tool for Exploring LLM Use in Group Ideation Tasks. In *IUI Workshops*. <https://hai-gen.github.io/2024/papers/1541-Gonzalez.pdf>

- Gordetzki, P., Blohm, I., & Hofstetter, R. (2023). Generative AI in Idea Development: The Role of Numeric and Visual Feedback. In *ICIS 2023 Proceedings*. <https://aisel.aisnet.org/icis2023/techandfow/techandfow/14>
- Goss-Sampson, M. A. (2022). *Statistical Analysis in JASP 0.16.1*. <https://jasp-stats.org/wp-content/uploads/2022/04/Statistical-Analysis-in-JASP-A-Students-Guide-v16.pdf>
- Gozzo, M., Woldendorp, M. K., & Rooij, A. de. (2022). Creative Collaboration with the “Brain” of a Search Engine: Effects on Cognitive Stimulation and Evaluation Apprehension. In M. Wölfel, J. Bernhardt, & S. Thiel (Eds.), *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. ArtsIT, Interactivity and Game Creation* (Vol. 422, pp. 209–223). Springer. https://doi.org/10.1007/978-3-030-95531-1_15
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology. Human Mental Workload* (Vol. 52, pp. 139–183). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2025). Complementarity in human-AI collaboration: concept, sources, and evidence. *European Journal of Information Systems*, 34(6), 979–1002. <https://doi.org/10.1080/0960085X.2025.2475962>
- Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. In *25th Pacific Asia Conference on Information Systems*, Virtual Event / Dubai, UAE. <https://aisel.aisnet.org/pacis2021/78>
- Holzner, N., Maier, S., & Feuerriegel, S. (2025). Generative AI and Creativity: A Systematic Literature Review and Meta-Analysis. <https://arxiv.org/abs/2505.17241v1>
- Houde, S., Brimijoin, K., Muller, M., Ross, S. I., Silva Moran, D. A., Gonzalez, G. E., Kunde, S., Foreman, M. A., & Weisz, J. D. (2025). Controlling AI Agent Participation in Group Conversations: A Human-Centered Approach. In *IUI '25 Proceedings*. ACM. <https://doi.org/10.1145/3708359.3712089>
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.
- Hwang, A. H.-C., & Won, A. S. (2021). IdeaBot: Investigating Social Facilitation in Human-Machine Team Creativity. In *CHI '21 Proceedings*. ACM. <https://doi.org/10.1145/3411764.3445270>
- JASP. (2023). *(Version 0.17)[Computer software]*.
- Javadi, E., Gebauer, J., & Mahoney, J. (2013). The Impact of User Interface Design on Idea Integration in Electronic Brainstorming: An Attention-Based View. *Journal of the*

- Association for Information Systems*, 14(1), 1–21.
<https://doi.org/10.17705/1jais.00322>
- Jiang, M., Karanasios, S., & Breidbach, C. (2024). Generative AI In The Wild: An Exploratory Case Study of Knowledge Workers. In *ECIS 2024 Proceedings*.
<https://aisel.aisnet.org/ecis2024/track04impactai/track04impactai/7>
- Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *J. Pers. Soc. Psychol.*, 44(1), 78–94.
<https://doi.org/10.1037/0022-3514.44.1.78>
- Kohn, N. W., & Smith, S. M. (2011). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 25(3), 359–371.
<https://doi.org/10.1002/acp.1699>
- Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1), 13601.
<https://doi.org/10.1038/s41598-023-40858-3>
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task.
<https://arxiv.org/abs/2506.08872v1>
- Krogh, G. von (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, 4(4), 404–409.
<https://doi.org/10.5465/amd.2018.0084>
- Kumar, H., Vincentius, J., Jordan, E., & Anderson, A. (2025). Human Creativity in the Age of LLMs: Randomized Experiments on Divergent and Convergent Thinking. In *CHI'25 Proceedings*.
- Lamm, H., & Trommsdorff, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. *European Journal of Social Psychology*, 3(4), 361–388. <https://doi.org/10.1002/ejsp.2420030402>
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.*, 37(6), 822–832.
<https://doi.org/10.1037/0022-3514.37.6.822>
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. *Organizational Research Methods*, 11(4), 815–852. <https://doi.org/10.1177/1094428106296642>
- Lee, B. C., & Chung, J. J. (2024). An empirical investigation of the impact of ChatGPT on creativity. *Nature Human Behaviour*, 8(10), 1906–1914.
<https://doi.org/10.1038/s41562-024-01953-1>
- Leggett Dugosh, K., & Paulus, P. B. (2005). Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology*, 41(3), 313–320.
<https://doi.org/10.1016/j.jesp.2004.05.009>

- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Liu, M., Ke, W., & Xu, D. J. (2023). Will Humans be Free-Riders? The Effects of Expectations for AI on Human-AI Team Performance. *PACIS 2023 Proceedings*.
- Maaravi, Y., Heller, B., Shoham, Y., Mohar, S., & Deutsch, B. (2021). Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science*, 15(6), 1431–1464. <https://doi.org/10.1007/s11846-020-00400-5>
- Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. <https://doi.org/10.1016/j.jbusres.2020.07.045>
- Mayring, P. (2014). *Qualitative Content Analysis - Theoretical Foundation, Basic Procedures and Software Solution*.
- McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1), 30–46. <https://doi.org/10.1037/1082-989X.1.1.30>
- Memmert, L. (2024). Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects. In *Proceedings of the 32nd European Conference on Information (ECIS)*, Paphos, Cyprus. https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1
- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/107289>
- Memmert, L., Cvetkovic, I., Tavanapour, N., & Bittner, E. (2025). Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load. *Business & Information Systems Engineering*. Advance online publication. <https://doi.org/10.1007/s12599-025-00974-y>
- Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. <https://aisel.aisnet.org/icis2024/aiinbus/aiinbus/7/>
- Memmert, L., Soroko, D., & Bittner, E. (2025). From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with

- Generative AI. *Business & Information Systems Engineering*(67), 615–635. <https://doi.org/10.1007/s12599-025-00960-4>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
- Michinov, N. (2012). Is Electronic Brainstorming or Brainwriting the Best Way to Improve Creative Performance in Groups? An Overlooked Comparison of Two Idea-Generation Techniques. *Journal of Applied Social Psychology*, 42, E222-E243. <https://doi.org/10.1111/j.1559-1816.2012.01024.x>
- Mirbabaie, M., Stieglitz, S., Brünker, F., Hofeditz, L., Ross, B., & Frick, N. R. J. (2021). Understanding Collaboration with Virtual Assistants – The Role of Social Identity and the Extended Self. *Business & Information Systems Engineering*, 63(1), 21–37. <https://doi.org/10.1007/s12599-020-00672-x>
- Muller, M., Houde, S., Gonzalez, G., Brimijoin, K., Ross, S. I., Moran, D. A. S., & Weisz, J. D. (2024). Group Brainstorming with an AI Agent: Creating and Selecting Ideas. In *International Conference on Computational Creativity*. https://computationalcreativity.net/iccc24/papers/ICCC24_paper_18.pdf
- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nijstad, B. A., Dreu, C. K. W. de, Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1), 34–77. <https://doi.org/10.1080/10463281003765323>
- NISO CRediT Working Group. *ANSI/NISO Z39.104-2022, CRediT, Contributor Roles Taxonomy*. NISO.
- Nomura, M., Ito, T., & Ding, S. (2024). Towards Collaborative Brainstorming among Humans and AI Agents: An Implementation of the IBIS-based Brainstorming Support System with Multiple AI Agents. In *ACM Collective Intelligence Conference Proceedings*. ACM. <https://doi.org/10.1145/3643562.3672609>
- OpenAI. (2023). <https://beta.openai.com/docs/api-reference/completions/create#completions/create-temperature>
- Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49. <https://doi.org/10.1016/j.tsc.2023.101356>
- Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.

- Paas, F., Renkl, A., & Sweller, J. [John] (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, 38(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1
- Paulus, P. (2000). Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied Psychology*, 49(2), 237–262. <https://doi.org/10.1111/1464-0597.00013>
- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Rafner, J., Beaty, R. E., Kaufman, J. C., Lubart, T., & Sherson, J. (2023). Creativity in the age of generative AI. *Nature Human Behaviour*, 7(11), 1836–1838. <https://doi.org/10.1038/s41562-023-01751-1>
- Redifer, J. L., Bae, C. L., & Zhao, Q. (2021). Self-efficacy and performance feedback: Impacts on cognitive load during creative thinking. *Learning and Instruction*, 71. <https://doi.org/10.1016/j.learninstruc.2020.101395>
- Reinig, B. A., Briggs, R. O., & Nunamaker, J. F. (2007). On the Measurement of Ideation Quality. *Journal of Management Information Systems*, 23(4), 143–161. <https://doi.org/10.2753/MIS0742-1222230407>
- Santanem, E. L., Briggs, R. O., & Vreede, G.-J. de (2004). Causal Relationships in Creative Problem Solving: Comparing Facilitation Interventions for Ideation. *Journal of Management Information Systems*, 20(4), 167–198. <https://doi.org/10.1080/07421222.2004.11045783>
- Sawyer, R. K. (2021). The iterative and improvisational nature of the creative process. *Journal of Creativity*, 31. <https://doi.org/10.1016/j.yjoc.2021.100002>
- Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models. *Business & Information Systems Engineering*, 66(2), 221–231. <https://doi.org/10.1007/s12599-024-00851-0>
- Siangliulue, P., Arnold, K. C., Gajos, K. Z., & Dow, S. P. (2015). Toward Collaborative Ideation at Scale. In *CSCW '15 Proceedings* (pp. 937–945). ACM. <https://doi.org/10.1145/2675133.2675239>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Siemon, D. (2022). Elaborating Team Roles for Artificial Intelligence-based Teammates in Human-AI Collaboration. *Group Decision and Negotiation*, 31(5), 871–912. <https://doi.org/10.1007/s10726-022-09792-z>
- Siemon, D. (2023). Let the computer evaluate your idea: evaluation apprehension in human-computer collaboration. *Behaviour & Information Technology*, 42(5), 459–477. <https://doi.org/10.1080/0144929X.2021.2023638>

- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.
- Siemon, D., & Wank, F. (2021). Collaboration With AI-Based Teammates – Evaluation of the Social Loafing Effect. In *PACIS 2021 Proceedings*. <https://aisel.aisnet.org/pacis2021/146/>
- Specht, A. R., & Buck, E. B. (2019). Crowdsourcing Change: An Analysis of Twitter Discourse on Food Waste and Reduction Strategies. *Journal of Applied Communications*, 103(2).
- Specker, R. J., Bucher, A., Katsiuba, D., Dolata, M., & Schwabe, G. (2025). An Extra Brain In The Room: Enhancing Large Group Brainstorming With An Autonomous AI-Agent. In *ECIS 2025 Proceedings*. https://aisel.aisnet.org/ecis2025/human_ai/human_ai/15/
- Stieglitz, S., Mirbabaie, M., Möllmann, N. R. J., & Rzycki, J. (2022). Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. *Information Systems Frontiers*, 24(3), 745–770. <https://doi.org/10.1007/s10796-021-10201-0>
- Summers-Stay, D., Voss, C. R., & Lukin, S. M. (2023). Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.
- Sweller, J [J.] (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., & Rintel, S. (2024). The Metacognitive Demands and Opportunities of Generative AI. In *CHI '24 Proceedings*.
- Tao, Y., Yoo, C., & Animesh, A. (2023). AI Plus Other Technologies? The Impact of ChatGPT and Creativity Support Systems on Individual Creativity. In *Proceedings of the 44th International Conference on Information Systems (ICIS)*. <https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/1>
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. <https://doi.org/10.1038/s41562-024-02024-1>
- Wang, K., & Nickerson, J. V. (2017). A literature review on individual creativity support systems. *Computers in Human Behavior*, 74, 139–151. <https://doi.org/10.1016/j.chb.2017.04.035>
- Wang, K., & Nickerson, J. V. (2019). A Wikipedia-based Method to Support Creative Idea Generation: The Role of Stimulus Relatedness. *Journal of Management Information Systems*, 36(4), 1284–1312. <https://doi.org/10.1080/07421222.2019.1661095>
- Wu, Z [Zhuohao], Ji, D., Yu, K., Zeng, X., Wu, D., & Shidujaman, M. (2021). AI Creativity and the Human-AI Co-creation Model. In M. Kurosu (Ed.), *Lecture Notes in*

- Computer Science. Human-Computer Interaction. Theory, Methods and Tools* (Vol. 12762, pp. 171–190). Springer International Publishing. https://doi.org/10.1007/978-3-030-78462-1_13
- Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *CHI '23 Proceedings*. ACM. <https://doi.org/10.1145/3544548.3581388>
- Zhan, X., Sun, H., & Fang, Y. (2024). The Janus-Faced Effects of Ideation with Generative AI on Individual Creativity and Dignity. In *ICIS 2024 Proceedings*. <https://aisel.aisnet.org/icis2024/userbehav/userbehav/23>
- Zhang, Z., Peng, W., Chen, X., Cao, L., & Li, T. J.-J. (2025). LADICA: A Large Shared Display Interface for Generative AI Cognitive Assistance in Co-located Team Collaboration. In *CHI'25 Proceedings*. ACM. <https://doi.org/10.1145/3706598.3713289>
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z [Zhanghao], Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2021). The effect of rank-ordering strategy on creative idea selection performance. *Eur. J. Soc. Psych.*, 51(2), 360–376. <https://doi.org/10.1002/ejsp.2743>

15.8 Appendix

15.8.1 Appendix A: Information Flow Overview

This screenshot (see Fig. A1) contains an overview of the interaction between the human and the brainstorming app as well as the information flow within the brainstorming app and towards the external generative language model (GLM) provider. Ideas entered by the user are displayed on the left. If the user clicks the button “Generate Ideas” (indicated by a green dot), a request is initiated to the external GLM provider, here, OpenAI’s GPT-3.5. The entire canvas contents of the participant, i.e., brainstorming question, list of ideas of the participant (if not empty), and list of previous GLM ideas for this participant (if not empty), are used to populate a predefined prompt template (see Appendix B). The GLM response to the prompt is split up into three separate ideas, and these are added as GLM ideas on top of the list of the “list of AI suggestions” on the right.

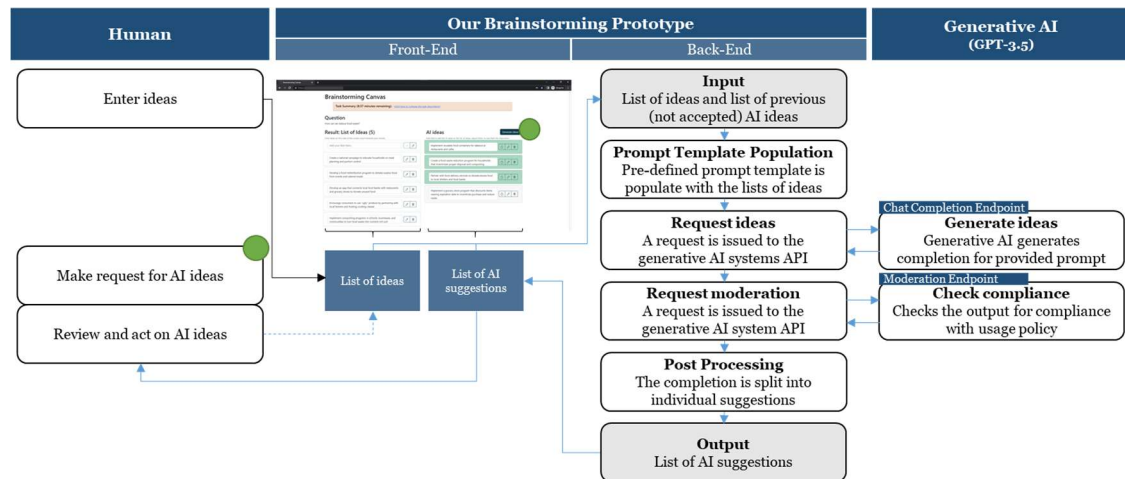


Fig. A.1 User interaction with and information flow within the brainstorming app

15.8.2 Appendix B: Prompt Template Development

The output of the GLM (in our study, the system-generated idea suggestions) is dependent on the prompt, making the prompt a critical component for the app (Schneider et al., 2024). The figure below contains the prompt template. The template is defined at development time and populated dynamically during runtime (see Appendix A for details). In this prompt template, curly brackets indicate placeholders, all caps placeholder labels (blue) indicate parameters fixed for this study, small caps placeholder labels (green) indicate parameters changing based on the current state of the canvas, and portions in italic font are only included if ideas are already present in the respective lists for the specific participant. We have adapted the prompt template from previous studies (blinded for review, blinded for review) and used the itemization prompt engineering technique (Mishra et al., 2022) when creating this template. This template is flexible to accept other brainstorming questions (see respective placeholder). The parameter values fixed for this study were:

- BRAINSTORMING_QUESTION= How can we reduce food waste?
- NUMBER_OF_IDEAS=3

Legend: ■ Parameters fixed for this study ■ Parameters changing based on current canvas state (only populated if present)

You are part of a brainstorming team. Your goal is to come up with novel and valuable ideas for the following question: {BRAINSTORMING_QUESTION}
Discarded ideas so far:
{list_glm_ideas from this human+GLM session}
Novel and valuable ideas so far:
{long_list_ideas from this human+GLM session}
 Please come up with {NUMBER_OF_IDEAS} additional novel and valuable ideas for the question: "{BRAINSTORMING_QUESTION}". Please provide the {NUMBER_OF_IDEAS} additional ideas as enumerated, ordered list. Each idea should be limited to a maximum of 20 words.

Fig. B.1 Prompt template with placeholders

15.8.3 Appendix C: Survey Instruments

Table C1. Survey Instruments

Concept	Reference	Asked to	Question/Text	Scale	Scale_Detail
De-mographics	---	ALL	How old are you?	nu- meric	
		ALL	What is your gender?	single select	
		ALL	What is you current study pro- gram?	single select	
Cognitive Load	Hart and Staveland (1988)	ALL	Mental Demand: How mentally demanding was the task?	21- point	Very low- Very high
		ALL	Physical Demand: How physi- cally demanding was the task?	21- point	Very low- Very high
		ALL	Temporal Demand: How hurried or rushed was the pace of the task?	21- point	Very low- Very high
		ALL	Performance: How successful were you in accomplishing what you were asked to do?	21- point	Perfect-Fail- ure
		ALL	Effort: How hard did you have to work to accomplish your level of performance?	21- point	Very low- Very high
		ALL	Frustration: How insecure, dis- couraged, irritated, stressed, and annoyed were you?	21- point	Very low- Very high
Cognitive Stimulation	Adapted from Gozzo et al. (2022)	HU- MAN+ AI	Please indicate to what extent the following statements apply to you (there is no right or wrong an- swers here).	-	
		HU- MAN+ AI	I felt inspired during the brain- storming task	5- point- likert	Strongly Agree- Strongly Dis- agree
		HU- MAN+ AI	The input of the AI did not help me	5- point- likert	Strongly Agree- Strongly Dis- agree

Concept	Reference	Asked to	Question/Text	Scale	Scale_Detail
Felt Responsibility	---	HU-MAN+ AI	The input of the AI helped me to think creatively	5-point-likert	Strongly Agree-Strongly Disagree
		HU-MAN+ AI	Working together with the AI at this moment was uninspiring	5-point-likert	Strongly Agree-Strongly Disagree
		HU-MAN+ AI	I could easily come up with associations	5-point-likert	Strongly Agree-Strongly Disagree
		HU-MAN+ AI	It was difficult to relate the input of the AI to my own thinking	5-point-likert	Strongly Agree-Strongly Disagree
	---	HU-MAN+ AI	For each result type described below, please indicate how you feel about your share of responsibility. Note: '0' means the AI was fully responsible...'100' means I feel I was fully responsible	-	
	---	HU-MAN+ AI	For the 'list of ideas' (prior to final idea selection) my share of responsibility was [%]:	numeric	0-100
	---	HU-MAN+ AI	For the 'list of 4 best ideas' (after the final idea selection) my share of responsibility was [%]:	numeric	0-100
	---	HU-MAN+ AI	Please briefly explain your reasoning for the previous answer for the share of responsibility regarding the result type 'list of ideas' (prior to final idea selection):	long free-text	
	---	HU-MAN+ AI	Please briefly explain your reasoning for the previous answer for the share of responsibility regarding the result type 'list of 4 best ideas' (after the final idea selection):	long free-text	
	---	HU-MAN+ AI	To achieve best performance when working with the tool with AI suggestions in a brainstorming task, I should...	5-point-likert	Focus on selecting and adjusting AI suggestions-Focus on generating ideas myself
GLM Effect	---	HU-MAN+ AI	How did the task change for you due to the presence of the AI (as compared to brainstorming alone)?	long free-text	
	---	HU-MAN+ AI	How did the AI (ideas/suggestions) help you in this brainstorming task?	long free-text	
	---	HU-MAN+ AI	How did the AI (ideas/suggestions) hinder you in this brainstorming task (i.e., made it difficult to perform the brainstorming task)?	long free-text	

15.8.4 Appendix D: Variable Operationalization

Table D.1 Operationalization of constructs (calculated per brainstorming session, considering only ideas of this session)

Con-struct	Variable name	List Basis	Operationalization	Rated by/ Data source	Reference(s)
Fluency	Fluency	long list	Number of ideas produced within the 10-minute session, according to the system log	system log	Paulus (2000)
Flexibility	Flexibility	long list	Number of distinct categories with ≥ 1 idea in the category system of Specht and Buck (2019)	blind student coder	Nijstad et al. (2010), Althuizen and Reichel (2016)
Novelty	Novelty	short list	Mean novelty on a 7-point Likert scale	5 blind-to-condition human judges	Siangliulue (2015), Siangliulue et al. (2015), Zhu et al. (2021)
	NoveltyGPT-rated	short list	Mean novelty on a 7-point Likert scale	GPT-4	Dell'Acqua et al. (2023), Haase and Hanel (2023), Organisciak et al. (2023)
	Good-Idea-Count- Novelty	long list	Number of long-list ideas with GPT-4 novelty rating $>$ avg (avg. novelty calculated across long lists from all sessions)	GPT-4	Reinig et al. (2007)
Value	Value	short list	Mean value-rating on a 7-point Likert scale	5 blind-to-condition human judges	Siangliulue (2015), Siangliulue et al. (2015), Zhu et al. (2021)
	ValueGPT-rated	short list	Mean value-rating on a 7-point Likert scale	GPT-4	Dell'Acqua et al. (2023), Haase and Hanel (2023), Organisciak et al. (2023)
	Good-Idea-Count- Value	long list	Number of long-list ideas with GPT-4 value-rating $>$ avg (avg. value calculated across long lists from all sessions)	GPT-4	Reinig et al. (2007)
Cognitive Load	NASA-TLX	—	NASA-TLX overall score (6-dim self-report)	participant self-report	Hart and Staveland (1988)

All idea-related variables (i.e., all variables except for cognitive load) were calculated on the level of each separate brainstorming session. Each variable was calculated on two performance levels (also see Fig. 2 in the main manuscript): human-level (considering only ideas of human origin) and human-GLM dyad-level (considering ideas of both human and GLM origin).

15.8.5 Appendix E: Coding Schema

Table E.1 Coding schema for analyzing the open-ended survey responses

Category	Code	Explanation & Example
Cognitive stimulation/inertia		Participant describes how the GLM affected their idea generation, e.g., stimulating or constraining it.
	stimulation	GLM helped initiate or expand ideas (e.g., new perspectives, entry support). “The ideas on the side which came from the AI were a great inspiration for developing new ideas on my own.” (P67)
	fixation	GLM ideas narrowed thinking or governed the ideation process. “It definitely felt like my thought process was being governed by the AI” (P48)
Effort (in-)dispensability		Participant reflects on the (non-)necessity, reduction, or redirection of their own effort.
	effort reduced	The participant relied on the GLM, putting in less effort themselves. “i became lazy” (P61)
	effort dispensable	Participant perceives that their own effort or own contribution is (somewhat) unnecessary (compared to the GLM’s contribution) for (parts of) the task. “i am not that creative” (P60)
	effort indispensable	The participant perceives their effort/ contribution as still essential. “The AI ideas seemed uncreative and boring” (P51)
	effort shift	The participant switched focus, e.g., from generating to curating or extending ideas. “All I had to do was read the ideas put forth and decide which ones make the cut.” (P48)
Responsibility (non-)diffusion		The participant provides their rationale for their perceived share of responsibility for the result.
	contribution-based	Rationale based on the origin of ideas, or share of ideas used or contributed (incl. copy/editing). “I selected 3 of my answers and 1 AI answer” (P54)
	curation responsibility	Rationale based on selection or evaluation effort for the ideas. “AI can give ideas, but I felt like the selection of the 4 ideas is the key point in concluding such [a] task.” (P52)
Cognitive load		Participant describes how mentally demanding or effortless the task felt.
	load reduction	The task felt easier or more comfortable. “The burden fell off my shoulders and instead of thinking hard, I let the AI write ideas.” (P46)
	mental disengagement	The participant mentally disengaged or stopped thinking. “I feel like using the AI made my brain shut off a little bit” (P47)
	load increase	Working with GLM required additional mental effort. “I had to reject the ai suggestions and pay attention not to be influenced too much by it or to be bogged down in the ai’s way of thinking.” (P75)

15.8.6 References

Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of*

- Management Information Systems*, 33(1), 11–44.
<https://doi.org/10.1080/07421222.2016.1172439>
- Dell’Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.
- Gozzo, M., Woldendorp, M. K., & Rooij, A. de. (2022). Creative Collaboration with the “Brain” of a Search Engine: Effects on Cognitive Stimulation and Evaluation Apprehension. In M. Wölfel, J. Bernhardt, & S. Thiel (Eds.), *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. ArtsIT, Interactivity and Game Creation* (Vol. 422, pp. 209–223). Springer. https://doi.org/10.1007/978-3-030-95531-1_15
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology. Human Mental Workload* (Vol. 52, pp. 139–183). Elsevier. [https://doi.org/10.1016/S0166-4115\(08\)62386-9](https://doi.org/10.1016/S0166-4115(08)62386-9)
- Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk’s Language. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 589–612).
- Nijstad, B. A., Dreu, C. K. W. de, Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1), 34–77. <https://doi.org/10.1080/10463281003765323>
- Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49. <https://doi.org/10.1016/j.tsc.2023.101356>
- Paulus, P. (2000). Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied Psychology*, 49(2), 237–262. <https://doi.org/10.1111/1464-0597.00013>
- Reinig, B. A., Briggs, R. O., & Nunamaker, J. F. (2007). On the Measurement of Ideation Quality. *Journal of Management Information Systems*, 23(4), 143–161. <https://doi.org/10.2753/MIS0742-1222230407>
- Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models. *Business & Information Systems Engineering*, 66(2), 221–231. <https://doi.org/10.1007/s12599-024-00851-0>

- Siangliulue, P., Arnold, K. C., Gajos, K. Z., & Dow, S. P. (2015). Toward Collaborative Ideation at Scale. In *CSCW '15 Proceedings* (pp. 937–945). ACM. <https://doi.org/10.1145/2675133.2675239>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Specht, A. R., & Buck, E. B. (2019). Crowdsourcing Change: An Analysis of Twitter Discourse on Food Waste and Reduction Strategies. *Journal of Applied Communications*, 103(2).
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2021). The effect of rank-ordering strategy on creative idea selection performance. *Eur. J. Soc. Psych.*, 51(2), 360–376. <https://doi.org/10.1002/ejsp.2743>

16 Publication 7: Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance

Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. <https://aisel.aisnet.org/icis2024/aainbus/aainbus/7/>

Abstract

Generative language models (GLMs) are used for creative tasks such as brainstorming. While studies show GLMs' creative potential, research on who benefits from support in creative tasks is inconclusive, and concerns are raised on whether humans reduce their effort when working with GLMs. Therefore, we explored how creative ability and GLM support affect the two brainstorming performance metrics fluency and originality. In our experiment with a GPT-4-based brainstorming app, we randomly assigned the 58 participants to one of two conditions: with or without GLM-support. Additionally, we split participants into a low- and a high-creativity group based on a creativity ability pre-test. Our findings reveal that low-creativity participants show increased fluency when being GLM-supported, while the effect on originality was positive for both groups. We contribute to the literature on human-GLM brainstorming and individual creativity support systems, offering a more nuanced perspective on GLMs' (potentially undesired) effects on behavior and performance.

Keywords: Generative Language Model, Human-AI Collaboration, Brainstorming, Creativity

16.1 Introduction

Generative language models (GLMs) like ChatGPT are increasingly being adopted in work settings even for creative work (Benbya et al., 2024). While traditionally, creativity was perceived to be a human strength when compared to AI (Dellermann et al., 2019), recent studies on idea generation showed that GLMs can produce ideas comparable to humans in quality (Haase & Hanel, 2023). However, the best ideas still originated from humans (Haase & Hanel, 2023; Koivisto & Grassini, 2023). Thus, instead of having humans and GLMs produce ideas separately from each other, it was proposed to explore joint human-GLM idea generation to better understand the effect on performance (Di Fede et al., 2022).

Several studies have since explored idea generation of humans together with GLMs in classical brainstorming settings, where the human contributes own ideas, but can also request suggestions from the GLM, generally finding a positive effect on different brainstorming performance metrics (Memmert & Bittner, 2024). However, if and how humans benefit from being supported by a technical system during ideation might be dependent on individual characteristics of the human such as the creative ability (Wang & Nickerson, 2017).

Literature on whether the less or the highly creative benefit more from non-AI-based individual creativity support systems (ICSS) is inconclusive (Wang & Nickerson, 2017). Given that humans can view AI systems as partners when working together (Rix, 2022), and given that GLMs can produce ideas comparable to humans (Haase & Hanel, 2023), recent literature suggests that humans might respond to GLMs similarly as to other human brainstorming group members (Liu et al., 2023). In purely human groups, humans might benefit from other's ideas, as these ideas can act as stimuli to develop additional, own ideas (Dugosh et al., 2000). However, when humans feel their effort to be more dispensable (i.e., less relevant or instrumental for group performance), they might reduce their effort, a behavior referred to a social loafing (Karau & Williams, 1993; Latané et al., 1979). It was suggested that such social loafing behavior might occur as well when working with GLMs (Liu et al., 2023). Given that social loafing behavior is influenced by factors including personal ability, we explore the following research question:

How does individual creative ability and brainstorming with a GLM affect brainstorming performance?

We conducted a between subject experiment with 58 participants. Participants performed a pre-test to determine their creativity level. Afterwards, they brainstormed for 8 minutes on a question related to climate change in a brainstorming app developed based on prior research (see Figure 1). The participants were randomly assigned to either brainstorming alone or to being supported by a GLM. After the brainstorming sessions, participants filled in a post-questionnaire regarding their experience. Performance was measured along the well-established brainstorming performance metrics fluency (i.e., number of ideas) and originality. For analysis, following Dell'Acqua et al. (2023), we split participants into low-creativity and high-creativity participants using the median pre-test creative ability test score. The randomly assigned condition and the split along creative ability score resulted in four similarly-sized groups. We conducted ANOVA analyses as well as exploratory statistical analyses to answer our research question.

We find that a higher fluency (higher number of ideas) was achieved only when low-creativity participants worked with the GLM, but that both low-creativity and high-creativity participants reached higher originality scores when working with GLM. However, working with the GLM comes at the cost of both low- and high-creativity humans themselves producing fewer ideas, contrary to findings in ICSS literature. We discuss the results with regard to usage behavior and potential implications for performance. With our study, we contribute to the literature around humans working with GLMs on creative tasks (Dell'Acqua et al., 2023; Di Fede et al., 2022) specifically, as well as mechanisms in human-AI groups more generally (Liu et al., 2023; Makarius et al., 2020; Siemon et al., 2015; Siemon & Wank, 2021). Additionally, we contribute to the long research history of ICSS, which has surfaced conflicting results on whether less or highly creative humans benefit most from ICSS (Wang & Nickerson, 2017), by offering a nuanced analysis of how working with a GLM affects humans and their performance depending on their creative ability. For

practice—while offering GLM-based support during ideation seems generally advisable—tool designers should make users aware of potentially undesirable side effects and counter-act these through a sociotechnical design.

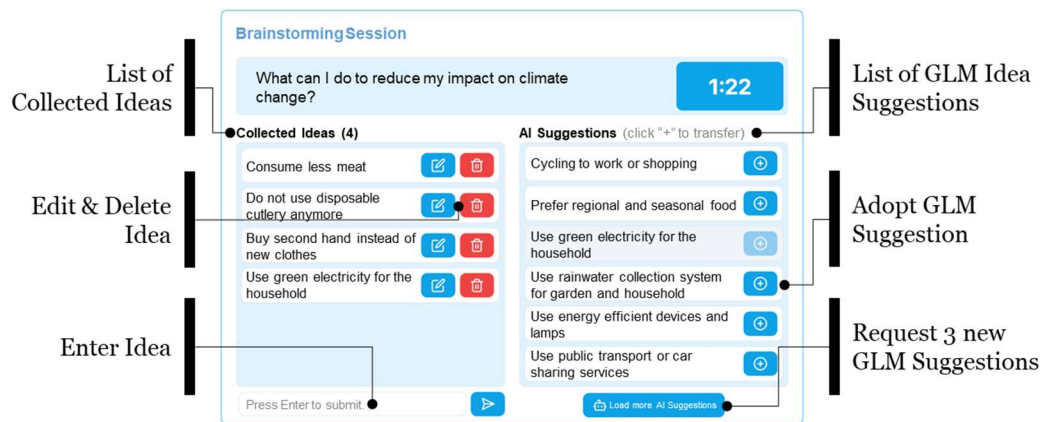


Figure 1 Translated screenshot of the GPT-4-based brainstorming app developed for this study (condition displayed: with GLM-support)

16.2 Background

16.2.1 Individual Creativity Support Systems (ICSS)

Creativity is a key driver for innovation and problem-solving, making it an important capability for individuals (Runco, 2004). In the context of creative idea generation, creativity refers to the ability to generate novel or original ideas, which are feasible, useful, and appropriate for the given problem (Runco & Jaeger, 2012). Individuals' creativity can be measured in several ways, e.g., by measuring the divergent thinking ability, i.e., the ability to generate various solutions to a problem (Olson et al., 2021).

Individual and group-level creativity are closely related (Mumford, 2012; Woodman et al., 1993). A popular technique for idea generation is *brainstorming* by Osborn (1953), who suggested four brainstorming rules. There is long research history on understanding how human groups can brainstorm effectively (Pinsonneault et al., 1999). Brainstorming performance is usually measured through fluency (i.e., number of ideas) and quality of ideas (Haase & Hanel, 2023; Nijstad et al., 2010; Paulus, 2000). Quality may include originality, novelty, or surprisingness as well value, or feasibility, though some of the concepts have been used interchangeably (Althuizen & Reichel, 2016; Diehl & Stroebe, 1987; Nijstad et al., 2010; Paulus, 2000).

When humans brainstorm in groups they can affect each other and their brainstorming performance through different *group mechanisms* (Pinsonneault et al., 1999), e.g., an idea of one group member might trigger a new idea for another group member (a mechanism referred to as *cognitive stimulation*), or, group members might reduce their effort when working with others (*social loafing*). Several tools were developed to support idea generation, both on a group-level and on an individual level (Wang & Nickerson, 2017), e.g., by in-

ducing or increasing the positive group mechanisms while removing or reducing the negative group mechanisms. Individual creativity support systems (ICSS) were developed to support individuals during idea generation (Althuizen & Reichel, 2016; Pinsonneault et al., 1999; Wang & Nickerson, 2017). A particularly effective type of ICSS are “stimuli provider” (Althuizen & Reichel, 2016), which offer certain cues to the human, which stimulate the human to search for additional ideas (Nijstad & Stroebe, 2006). Studies show that human fluency decreases over time during brainstorming sessions, i.e., humans run out of ideas (Baruah & Paulus, 2016; Kohn & Smith, 2011). When working with a stimuli provider ICSS, humans can request cues, e.g., when running out of ideas (Siangliulue, Chan, et al., 2015), to help them to produce additional ideas. While such systems are effective in supporting individual idea generation (Althuizen & Reichel, 2016), they have (technical and practical) limitations. Many studies used artificial settings with stimuli previously curated by the researchers (Dugosh et al., 2000), e.g., extracted from word association dictionaries (Althuizen & Reichel, 2016), or relied on social media to contain relevant stimuli (Siemon et al., 2015). Recent advancements of GLMs might offer the opportunity to overcome these limitations and improve on stimuli-provider ICSS. First studies show that GLMs can produce complete, creative ideas comparable to many humans (Haase & Hanel, 2023), without requiring prior task- or topic-specific training. We, therefore, explore joint brainstorming of humans with GLM-based systems

16.2.2 Generative Language Models (GLMs) and Brainstorming

Creativity is traditionally considered a comparative human strength compared to AI (Dellermann et al., 2019). However, GLMs were used for story writing, design science research, or innovation (Gero, 2023; Memmert et al., 2023; Steiger et al., 2024), showcasing their creative potential. When generating ideas in isolation (not together with humans), GLMs can produce creative ideas comparable to many humans, though the best ideas still originate from humans (Haase & Hanel, 2023; Organisciak et al., 2023). It was suggested, that instead of investigating humans and GLMs working in isolation, researchers should explore how humans and GLMs work together (Di Fede et al., 2022; Liu et al., 2023). Therefore, in this study, we explore humans using a GLM-based ICSS. Following classical ICSS brainstorming study settings, individual humans brainstorming with an app, but have the option to request (pull) ideas from the GLM when desired.

GLM-based ICSS might not only offer the opportunity to overcome the limitations of traditional ICSS (in requiring no prior curation of stimuli, or relying on readily expressed ideas), but also offer new interaction opportunities, which might make them fundamentally different from traditional ICSS. GLMs can produce full ideas at rapid pace, offering the user the choice to use them as inspiration, edit them, or to copy them (Memmert & Tavapour, 2023). Thus, GLM-based ICSS are similar to a traditional ICSS in that they provide external stimuli, but different, in that they can produce complete, human quality ideas. Gabriel et al. (2016) referred to ICSS capable to contribute ideas as a “colleague”, drawing a parallel to humans. Prior studies discussed that group mechanisms like in human groups

might arise in human-GLM dyads, e.g., GLM ideas might stimulate humans or humans might loaf (Liu et al., 2023; Memmert & Bittner, 2024; Memmert & Tavanapour, 2023), but it remains unclear how human and collective performance is affected.

Individual differences influence creativity (also see Mumford, 2012; Woodman et al., 1993) and might affect whether and how humans benefit from using ICSS. However, as Wang and Nickerson (2017) note, literature remains inconclusive on whether the less or the more creative benefit more. With the emergence of GLMs, a similar question has arisen regarding how individual differences, such as task ability, affect benefitting from GLM usage in creative tasks (Liu et al., 2023). Since the question of how the task ability affects who benefits from using ICSS or GLMs for creative tasks was raised in both research streams—ICSS literature (Wang & Nickerson, 2017), and human-GLM literature (Liu et al., 2023)—we suggest to investigate this further.

16.3 Method

16.3.1 System and Usage Process: Brainstorming App with GLM-Capability

To conduct our experiment, we set up a simple, web-based brainstorming app (see Figure 1) based on prior research (Di Fede et al., 2022; Memmert et al., 2024; Memmert & Tavanapour, 2023). The app displays a brainstorming question. Users can add, edit, and delete ideas from the list of “collected ideas” on the left. The right side contains a list of AI suggestions. The list is empty initially. Participants can make requests for AI suggestions through a button. Per request, three new suggestions are generated in the backend and added to the list of AI suggestions. AI suggestions can be copied to the list of ideas (left) individually and, if desired, edited. The AI suggestions were not (automatically) pushed to the participants, but displayed only on request, as this is most effective (Siangliulue, Chan, et al., 2015). We conducted four pretesting sessions.

Table 1 Prompt for brainstorming prototype

Purpose	Role	Prompt/Prompt Templates (placeholders in square brackets)
Initial context message	System	You are an expert in the domain of climate change and part of a brainstorming team. Generate three original and creative ideas related to <i>[topic]</i> . Each idea should be unique and not mentioned by you or the user before. Keep the ideas concise and in about 20 words. While considering the user's input, aim to explore new directions and encourage innovative thinking. Provide ideas in German.
New idea	User	Idea: <i>[idea content]</i>
Request suggestion	Assistant	Generate three new ideas.

The interaction with the GLM is hidden by the graphical user interface. When the user requests new AI suggestions, the app automatically builds a prompt, communicates with the GLM, and displays the suggestions. The app creates one conversation (stream of messages) per brainstorming session—hidden to the user. The initial context message provides

the necessary context information to the model. It was developed based on prompt engineering literature (Mishra et al., 2022) and based on similar apps (Memmert & Bittner, 2024). User ideas were added to the conversation as new messages automatically. When the user request AI suggestions, a message was added to instruct the GLM to generate three new ideas. The messages can be seen in Table 1. We selected GPT-4 as the underlying GLM as it is widely adopted and due to its high idea generation performance across topics (Haase & Hanel, 2023). Neither model nor the prompt were tailored towards the brainstorming topic of climate change (i.e., topic placeholder can be filled dynamically) and default settings were used. All necessary instructions regarding the task, topic, and expected output were delivered via the prompts that were used for communication with the model.

16.3.2 Hypothesis Derivation

Our research goal is to understand how the creative ability of humans affects if and how humans benefit from brainstorming together with a GLM. Along this research goal, we investigate how the creative ability, the presence of GLM-support and the interaction of these two factors affects performance, measured via the brainstorming performance metrics *fluency* and *originality* (Haase & Hanel, 2023). We thus investigate three (creative ability, GLM-support, interaction effect) pairs (fluency, originality) of hypotheses.

Divergent thinking like generating ideas for a problem is an important part of creative work (Olson et al., 2021; Wang & Nickerson, 2017). Several theories existing to explain creativity in general and the development of creative ideas in particular (Wang & Nickerson, 2017). A higher creative ability is associated with higher performance in ideation tasks including both producing more and more original ideas (Nijstad et al., 2010). Additionally, there seems to be a positive relationship between the quantity and quality of ideas (Nijstad et al., 2010). This assumption is reflected in the brainstorming rules (Osborn, 1953) which encourages producing many ideas, and adjusting and recombining ideas. We therefore hypothesize:

Hypothesis 1a: Participants with a high creativity level will achieve a significantly higher fluency (i.e., number of ideas) compared to participants with a low creativity level.

Hypothesis 1b: Participants with a high creativity level will achieve a significantly higher originality score compared to participants with a low creativity level.

Humans produce less and less ideas the longer they brainstorm, i.e., fluency decreases over time, as humans run out of ideas or get stuck thinking about a problem aspect (Baruah & Paulus, 2016; Kohn & Smith, 2011). External stimuli can activate relevant knowledge in human's memory and thereby help them to come up with new ideas (Althuizen & Reichel, 2016; Nijstad et al., 2010; Nijstad & Stroebe, 2006). Such stimuli could be provided in the form of ideas by other brainstorming group members or by ICSS, triggering the performance-enhancing effect of cognitive stimulation (Pinsonneault et al., 1999). Stimuli-pro-

vider ICSS help humans to develop new ideas by providing relevant cues such as brainstorming question-related words (Althuizen & Reichel, 2016). While earlier stimuli-provider ICSS typically used (groups of) words, GLMs can generate complete ideas of a quality comparable to humans. We suggest that these might act as (external) stimuli to support humans in generating additional ideas, e.g., when running out of own ideas.

Additionally, given that GLMs produce ideas of a quality comparable to many humans (Haase & Hanel, 2023), we expect participants to adopt (copy and/or adjust) many of the suggested GLM ideas, increasing fluency by adding the idea to the collection of ideas. Both, the assumed (indirect) positive influence on participant's idea generation (through cognitive stimulation) and the direct effect of the possibility of adopting GLM ideas suggest GLM-supported sessions might show higher fluency:

Hypothesis 2a: Participants with GLM support will achieve a significantly higher fluency (i.e., number of ideas) compared to participants without GLM support.

Prior research suggests that the most creative ideas are still produced by humans, not GLMs (Dell'Acqua et al., 2023; Haase & Hanel, 2023; Koivisto & Grassini, 2023). However, GLMs can help humans to cover the brainstorming question more holistically by developing more ideas in different categories related to the problem, i.e., increase flexibility (Memmert & Bittner, 2024). Increased flexibility is associated with higher originality (Althuizen & Reichel, 2016; Nijstad et al., 2010). Along with the positive relationship of fluency and originally discussed above, we therefore suggest:

Hypothesis 2b: Participants with GLM support will achieve a significantly higher originality score compared to participants without GLM support.

Lastly, we seek to understand if the creativity level and the GLM support show an interaction effect. From human groups it is known that humans calibrate their effort based on the performance of other group members, e.g., when working together with a more capable group member, they might reduce their effort, because they feel their own effort is dispensable. This effect of reducing one's effort in group settings is referred to as social loafing (Karau & Williams, 1993; Latané et al., 1979). The effect is well-documented in purely human groups. Additionally, several studies explored this effect in human-AI settings (Siemon et al., 2015; Siemon & Wank, 2021; Stieglitz et al., 2022). It was suggested to further explore this effect when individual humans work with GLMs (Liu et al., 2023; Memmert, 2024).

As discussed, GLMs can produce ideas of comparable quality to many humans (Haase & Hanel, 2023). We expect that participants will compare their own ideas to the GLM suggestions. Thus, one could expect, that humans who are less creative will produce fewer own ideas (Karau & Williams, 1993; Liu et al., 2023), because they feel their effort to be dispensable (fewer own ideas, i.e., lower fluency). Instead of developing own ideas, they might be more likely to adopt GLM ideas (more GLM ideas adopted, i.e., higher fluency). Thus, the effect on fluency is unclear. In contrast, more creative humans might be more critical when judging the quality of GLM ideas compared to their own ideas and might

consequently adopt fewer GLM ideas. However, in humans groups, stronger group members might compensate for less capable group members, a phenomenon referred to as social compensation (Williams & Karau, 1991). One might expect creative humans do *not* feel their effort to be dispensable, i.e., they might feel instrumental for the task outcome, and to thus invest more effort into developing own ideas. Thus, it is unclear, how creativity and GLM support will affect performance dimensions. In ICSS literature the question of “who do creativity support systems benefit more: the highly creative or the less creative?” surfaced mixed results (Wang & Nickerson, 2017, p. 147). We therefore suggest the following, undirected hypotheses:

Hypothesis 3a: There will be a significant interaction effect between creativity level and GPT support on fluency (i.e., number of ideas), indicating that the effect of the GPT support varies depending on the creativity level.

Hypothesis 3b: There will be a significant interaction effect between creativity level and GPT support on the originality score, indicating that the effect of the GPT support varies depending on the creativity level.

16.3.3 Procedure & Participants

We conducted an online experiment to test our hypotheses. The study started with a general introduction and the collection of basic demographic information. Participants performed a creative task without time limit to measure their creative ability (see measurement section for a detailed description). Afterwards, participants were presented with the brainstorming prototype (see Figure 1) and asked to brainstorm for 8 minutes. Brainstorming studies frequently use societal relevant questions (Memmert & Bittner, 2024; Zhu et al., 2021). For our study we selected the question of “What can I do to reduce my influence on climate change?”. Once the brainstorming session concluded, participants were presented with their “collection of ideas” and asked to select their four most original ideas, following the approach of Siangliulue, Arnold, et al. (2015). For selection, participants were presented with the same definition of originality that was later used by the expert to rate the ideas. In a post-questionnaire, we collected information regarding the experience. Study participation was voluntary and took about 20 minutes. Of the 77 participants who enrolled in the study, 58 completed study and were included in the analysis: 24 participants identified as female and 34 as male with ages ranging from 22 to 66, with an average of 30 years.

16.3.4 Treatment & Measurement

Below we describe the two independent variables (creative ability, GLM support) and the two dependent variables (fluency, originality score). Creative ability is a central construct for our study. For creative ability a variety of measurement instruments exist. We selected the Divergent Association Task (DAT), as it can be completed in a short amount of time and shows strong correlation with other, established creativity measurement instruments such as the Alternative Uses Task (Olson et al., 2021). In the DAT, participants are asked to come up with ten nouns that are as different from each other as possible in all meanings and uses of the words, independently from any specific topic or question (i.e., unrelated to

the brainstorming question). The DAT then uses the GloVe algorithm to calculate semantic distances between the words. Semantically similar words that tend to co-occur in text will be closer in the vector space, resulting in lower scores, e.g., the combination *arm* and *leg* would receive a lower score than *violin* and *tomato* (Olson et al., 2021, p. 2). The DAT scores is calculated as the average distance between the words of the participant.

We had participants randomly assigned to either the GLM support or the no-GLM support condition. Participants in the GLM support conditions were presented with the app as shown in Figure 1. They could request and adopt GLM suggestions, if desired. For participants in the no-GLM support condition, the option to request GLM suggestions was removed, i.e., participants had to brainstorm alone without a GLM.

To measure brainstorming performance, we selected the two well-established brainstorming performance measures fluency and originality (Haase & Hanel, 2023). Fluency refers to the number of ideas produced in a brainstorming session. Fluency was recorded on group level, i.e., included all ideas in the “idea collection”, regardless of whether they originated from the human participant or the GLM (adopted ideas). While fluency is a well-established metric in brainstorming literature, Haase and Hanel (2023) pointed out that GLMs can quickly produce a vast amounts of text, making it less meaningful. However, given that we do not compare humans and GLMs in isolation, given its importance in brainstorming literature, and given its positive correlation with originality (Nijstad et al., 2010), we report the metric here for completeness.

For originality, we had an expert, blind-to-condition judge rate all ideas that were selected as the most original ideas by participants. The judge is an expert in the domain of the brainstorming topic (climate change), working as an Associate Program Officer at the United Nations Framework Convention on Climate Change (UNFCCC) secretariat. The UNFCCC secretariat is the United Nations entity tasked with supporting the global response to the threat of climate change. The expert was provided with the same definition of originality as the participants. The following definition was used: “Originality is a critical factor in evaluating creative outcomes. As per established definitions in literature, originality refers to something that is unusual, new or unique. If something is not new or unusual, it is banal, conventional or commonplace - not original and therefore not creative. Commonly used synonyms for originality are: novelty, infrequency” (Runco & Jaeger, 2012). Following similar studies, originality of the ideas was rated on a Likert scale from 1 (not at all original) to 7 (very original). To ensure an unbiased rating, the ideas were pre-processed such that the expert had no information about whether an idea was generated by a human alone or with GLM assistance. This pre-processing involved correcting spelling (capitalization, incorrect spellings, commas, and sentence structure), standardizing formulations to avoid colloquial language and randomizing the presentation order for rating.

16.3.5 Analysis

We were interested in how creative ability and GLM support affect performance. Participants were assigned randomly to either the GLM support or no-GLM support condition, resulting in two equally-sized groups. We measured participants' creative ability using the DAT. DAT scores ranged between 61.98 and 93.19 (mean=81.38). Olson et al. (2021, p. 2), who proposed the DAT, report that DAT scores "commonly range from 65 to 90 and almost never exceed 100" and report a mean DAT score of 78.38 (SD = 6.35), indicating a similar level of creative ability. Following Dell'Acqua et al. (2023), for the analysis, we split participants into the two equally-sized groups of low-creativity and high-creativity using the median pre-test DAT score of 81.38. This resulted in the 58 participants being split in four, similarly-sized groups: low-creativity/no-GLM-support (n=15), low-creativity/GLM-support (n=14), high-creativity/no-GLM-support (n=14), high-creativity/GLM-support (n=15). For analysis, we performed the required assumption checks and conducted two-way ANOVAs for each of the two dependent variables (fluency, originality score) as well as post-hoc tests to understand group differences. Additionally, we conducted an exploratory analysis, using T-tests and correlation analysis to better understand the results. We also used the log-data of the interaction of the participants with the tool and visualized it to explore the participants' usage behavior of the app.

16.4 Results

To gain insights into the participants behavior during the brainstorming session and their interaction with the app, we recorded when the participants created own ideas, requested GLM suggestions, or adopted GLM suggestions (with or without adjustment). An overview regarding the events in the brainstorming sessions is presented in Figure 2. The 58 participants in total generated 1104 ideas, averaging 19 ideas per brainstorming session. 772 (69.93%) ideas were generated by humans alone, 320 (28.99%) were adopted GLM suggestions and 12 (1.09%) were based on GLM suggestions but edited by the participant. GLM-supported participants received 788 idea suggestions by requesting GLM suggestions 264 times (three suggestions per request). On average, each GLM-supported participant was presented with 27 GLM idea suggestions. Out of the 788 GLM suggestions presented, 332 were adopted (adoption share: 42.13%).

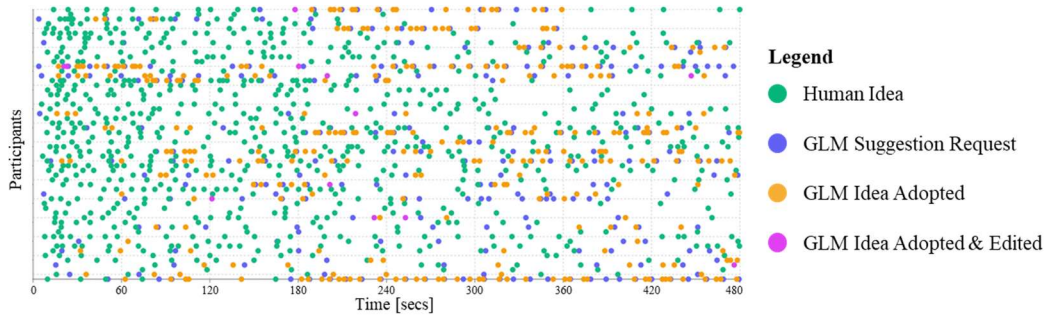


Figure 2 Events in participants' brainstorming sessions

16.4.1 Group Fluency (Number of Ideas)

We first report the results of the effects of creativity and GLM support (and their interaction) on fluency, i.e., number of ideas. For analysis, we performed a two-way ANOVA. The mean values and standard deviations for fluency are presented in Table 2. Given that the Levene's Test for equality of variances was significant ($p=.007$), results of the ANOVA analysis need to be interpreted with caution.

Table 2 Descriptive statistics for group fluency
(i.e., number of ideas including GLM ideas)

Creativity Level	GPT Support	Mean	Standard Deviation
Low	No	16.533	6.379
	Yes	28.000	11.529
High	No	16.071	6.044
	Yes	15.933	7.1850

We had hypothesized that high-creativity humans will show higher fluency (Hypothesis 1a). The results indicate a significant main effect of creativity level ($F(1, 54) = 8.790$, $p=.005$, partial $\eta=.140$) on fluency. However, contrary to our hypothesis, low-creativity participants (mean=22.069) produced more ideas compared to high-creativity participants (mean=16.000). Thus, hypothesis 1a is not supported. We expected GLM-supported humans to show higher fluency (Hypothesis 2a). The analysis shows a significant main effect of GPT support on fluency with medium to large effect size ($F(1, 54)=7.186$, $p=.010$, partial $\eta^2=.117$). GLM-supported humans (mean=21.759) on average produce more ideas compared to humans working without GLM-support (mean=16.310), i.e., hypothesis 2a is supported. Lastly, we hypothesized there to be an interaction effect of creativity level and GLM-support on fluency (Hypothesis 3a). The analysis revealed a significant interaction effect ($F(1, 54)=7.541$, $p=.008$, partial $\eta^2=.123$) on fluency. Thus, hypothesis 3a is supported, indicating low-creativity and high-creativity humans are affected differently by the GLM support. The pronounced effect of GLM support on participants with low creativity is depicted in the Figure 3, which shows that for participants with a low creativity level (represented by the red line), the presence of GPT support resulted in a substantial increase in the estimated marginal mean of fluency compared to participants without GLM support. A Tukey post-hoc test showed this difference to be significant with a large effect size ($p=.001$, Cohen's $d=1.483$). This suggests that GPT support provided a significant boost to

the number of ideas generated of low creativity participants. On the other hand, for participants with a high creativity level (represented by the blue line), the effect of GPT support was not statistically significant (Tukey post-hoc test: $p=1.000$).

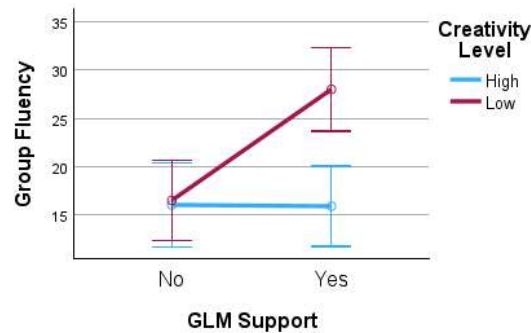


Figure 3 Effect of GLM support and creativity level on group fluency (i.e., number of ideas including GLM ideas)

16.4.2 Originality Score

Following the same procedure, we performed a two-way ANOVA for creativity level and GPT support on the originality score. The mean values and standard deviations for originality score are shown in Table 3.

Table 3 Descriptive statistics for originality score

Creativity Level	GLM Support	Mean	Standard Deviation
Low	No	2.883	0.550
	Yes	3.161	0.669
High	No	3.000	0.449
	Yes	3.517	0.495

We had hypothesized, that high creativity participants receive a higher originality score (Hypothesis 1b). However, the main effect of creativity level on originality score was not significant ($F(1, 54)=2.710$, $p=.106$), i.e., hypothesis 1b is not supported. We further expected GLM-supported participants to achieve a higher originality score (Hypothesis 2b). We find a significant main effect of GPT support on originality score ($F(1, 54)=7.649$, $p=.008$, partial $\eta^2=.124$); participants with low creativity received an average originality score of 2.940, compared to 3.345 for high-creativity participants. Thus, Hypothesis 2b is supported. Lastly, we had hypothesized that low-creativity and high-creativity participants are affected differently from GPT support (Hypothesis 3b). However, the interaction was not significant ($F(1, 54) = 0.695$, $p = .408$), i.e., hypothesis 3b is not supported. The results are visualized in Figure 4. The lines for low and high creativity levels show a similar slope, indicating a similar positive effect of GPT support on originality scores regardless of creativity level. A pairwise group comparison via Tukey post hoc test showed the low-creativity, non-GPT support group to be significantly different from the high-creativity, GPT-support group ($p=.013$). The difference between GPT-supported and non-GPT-supported high-creativity participants was marginally significant ($p=.064$). Table 4 summarizes the results for the supported and rejected hypotheses.

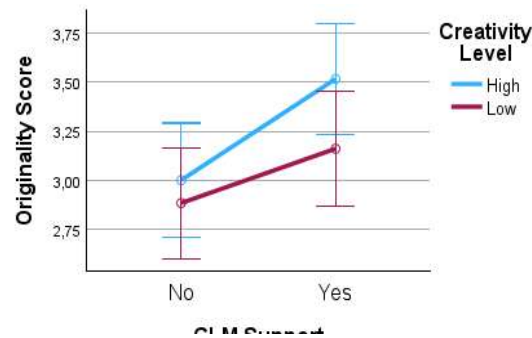


Figure 4 Effect of GPT support and creativity level on originality score

Table 4. Supported and rejected hypotheses

#	Hypothesis	Supported	Comment
H1a	Participants with a high creativity level will achieve a significantly higher fluency (i.e., number of ideas) compared to participants with a low creativity level.	No	Reversed
H1b	Participants with a high creativity level will achieve a significantly higher originality score compared to participants with a low creativity level.	No	Marginally significant (p=.106)
H2a	Participants with GPT support will achieve a significantly higher group fluency (i.e., number of ideas) compared to participants without GPT support.	Yes	-
H2b	Participants with GPT support will achieve a significantly higher originality score compared to participants without GPT support.	Yes	-
H3a	There will be a significant interaction effect between creativity level and GPT support on group fluency (i.e., number of ideas), indicating that the effect of the GPT support varies depending on the creativity level.	Yes	-
H3b	There will be a significant interaction effect between creativity level and GPT support on the originality score, indicating that the effect of the GPT support varies depending on the creativity level.	No	-

16.4.3 Exploratory analysis

To better understand the how working with the tool affects brainstorming performance, we conducted several exploratory follow-up analyses.

16.4.3.1 Group Fluency & Human Fluency over Time

As discussed, more ideas were produced (higher fluency) in GPT-supported sessions. We plotted the ideas produced over time. As expected from literature, fluency decreased for participants over time. However, for GLM-supported participants, the fluency decreased less compared to non-GLM supported participants (Figure 5a). When plotting the ideas produced in the GLM-supported brainstorming sessions differentiated by origin (Figure 5b), we find that humans first produce more own ideas, but later adopt more AI ideas.

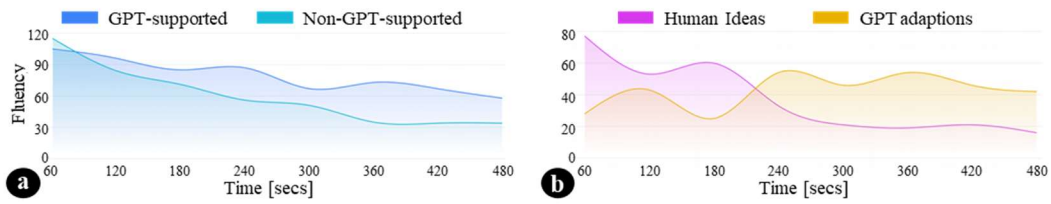


Figure 5 Produced ideas over time in a) GPT- and Non-GPT-supported brainstorming sessions, b) GPT-supported brainstorming session by origin (Human vs. AI)

16.4.3.2 Human Fluency

In Figure 5b, a drop in human ideas and an increase in AI adoptions occur. While for the main analysis (H1a, H2a, H3a) we used a collective perspective onto fluency (i.e., including both ideas of human and of AI origin, referred to as group fluency), we wanted to explore how working with the GLM affected the fluency of the human. While for the no-GLM-support condition necessarily all ideas originated from the human, for the GLM-support condition, ideas from the GLM were excluded. Table 5 shows the results.

Table 5 Descriptive statistics for human fluency (i.e., number of ideas excluding GLM ideas)

Creativity Level	GLM Support	Mean	Standard Deviation
Low	No	16.533	6.379
	Yes	11.071	5.863
High	No	16.071	6.044
	Yes	9.600	5.792

The main effect of creativity level on human fluency was not significant ($F(1, 54)=0.373$, $p=.544$, partial $\eta^2=.007$). However, we find a significant main effect of GPT-support on human fluency ($F(1, 54)=14.197$, $p<.001$, partial $\eta^2=.208$). GLM-supported humans (mean=10.310) produced significantly fewer ideas compared non-GLM supported humans (mean=16.310). The interaction effect of creativity-level and GPT-support on human fluency was not significant ($F(1, 54)=0.102$, $p=.751$, partial $\eta^2=.002$). Results indicate a reduction in human fluency when working with the app independently of creativity level (Figure 6). A pairwise group comparison via Tukey post hoc test revealed that high-creativity humans who were not-GLM-supported produced significantly more ideas compared to high-creativity humans with GLM-support ($p=0.028$).

16.4.3.3 Usage Behavior in GLM-Supported Groups Based on Creativity-Level

As shown in Figure 5b, humans working with GLMs produce fewer ideas themselves and start adopting more GLM suggestions over time. We explored if there is a difference in usage behavior depending the creativity-level for GLM-supported participants. In a first step, we looked at the number of requests for suggestions by the participants. As the number of suggestion requests was not normally distributed for the high-creativity group (Shapiro-Wilk: $W=0.812$, $p=.007$), we calculated the non-parametric Mann-Whitney-U test, which showed a significant ($U=153.0$, $p=.038$) difference with medium to large effect size (rank-biserial correlation $r_B=0.457$; see Goss-Sampson, 2022) between the number of suggestion requests for low-creativity ($mdn=9.5$) and high-creativity ($mdn=6$) participants. In second

step, we calculated the adoption share, as the number of ideas adopted (copied by the participants with or without adjustments) divided by the number of suggestions provided. The calculated Student's T-test shows a marginally significant difference ($t(27)=1.928$, $p=.064$) between the adoption share of low-creativity (mean=47.5%) and high-creativity (mean=31.6%) humans. Thus, low-creativity humans made more requests for suggestions, and (in tendency) showed a higher adoption share for the suggestions provided to them. This might explain the fluency (number of ideas) of GLM-supported low-creativity participants. Such intense use of the GLM could be interpreted as a reduction of own effort.

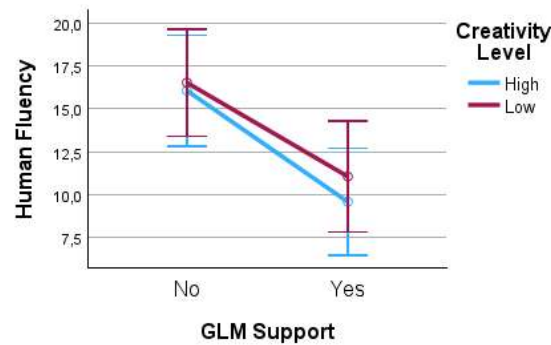


Figure 6 Effect of GLM support and creativity level on human fluency (i.e., number of ideas excluding GLM ideas)

16.4.3.4 Fluency and Originality

There seems to be a positive relationship between fluency and originality (Nijstad et al., 2010), potentially, as the most original ideas arise later in the brainstorming session, when the most obvious ideas are expressed and recombinations and adjustments of ideas lead to the more original ideas. Indeed, we find such a positive, significant correlation between group fluency and originality score for our study (Spearman's $\rho=.248$, $p<.031$), but not for human fluency (excluding GLM ideas) and originality score (Pearson's $r=-0.003$, $p=.985$).

16.4.3.5 Originality Ratings of GLM and Human Ideas

Given participants across creativity levels accepted many of the GLM suggestions and selected many of them as their most original ideas, we further explored idea originality, differentiating between ideas of human and GLM origin. Of the 232 ideas selected as most original by the participants (four per participant), 176 were of human and 56 of GLM origin. The normality assumption was not fulfilled for either set of ideas (Shapiro-Wilk – human: $W=0.866$, $p<.001$; GLM: $W=0.874$, $p<.001$). The calculated Mann-Whitney-U test shows a significant difference ($U=5861.0$, $p=.026$) with small to medium effect size ($r_B=0.189$) between human ideas ($mdn=3$, mean=3.034) and GLM ideas ($mdn=3$, mean=3.482) on originality ratings.

16.5 Discussion

16.5.1 General Discussion

Overall, the findings of how creative ability and GLM support affect brainstorming performance are mixed. Taking a collective, group outcome perspective on the brainstorming performance metric fluency, low-creativity humans benefit from GLM support, whereas high-creativity humans do not benefit. For originality, both low- and high-creativity humans showed increased performance. However, when taking an individual performance perspective, i.e., considering only ideas created by the human and excluding ideas of GLM origin, both low- and high-creativity humans produced fewer ideas themselves. There seems to be a difference in how low-creativity and high-creativity humans use the system, with low-creativity humans requesting and accepting more GLM suggestions compared to high-creativity humans. We discuss the results in more detail below with regard to the superior performance of human-AI groups, differences in usage behavior and potential explanations, potentially unintended side effects of introducing a GLM, GLMs in the context of stimuli-provider ICSS, and performance-affecting group mechanisms.

A key goal of human-AI collaboration is for the human-AI team to achieve superior performance compared to what the individual could have achieved alone (Dellermann et al., 2019). As performance is task-dependent, we choose fluency and originality as metrics, as these are well-established brainstorming performance metrics (Haase & Hanel, 2023). Indeed, we find the human-AI dyads to perform significantly better along both of those metrics compared to unaided individuals. This might have been surprising, given that idea generation is a creative task and creativity is usually considered to be a comparative strength of humans when compared to AI systems (Dellermann et al., 2019). However, our findings are consistent with recent literature showing the potential of using GLMs in creative tasks (Haase & Hanel, 2023; Organisciak et al., 2023). Our exploratory analysis showed that the GLM ideas on average were rated significantly more original than human ideas, making them a driver for performance improvement. This aligns with recent literature on the ability of GLMs to produce creative ideas comparable to humans (Haase & Hanel, 2023).

It is an ongoing area of research to understand how individual differences (such as task or creative ability) affect how humans use and are affected by GLMs (Dell'Acqua et al., 2023; Liu et al., 2023). We find low-creativity humans to show significantly better performance for both fluency and originality score when GLM-supported compared to low-creativity humans working alone, whereas the same applied only for originality for high-creativity humans. This seems to be consistent with recent findings in literature showing bottom-performers benefitting more from using a GLM on knowledge work and creative tasks specifically (Dell'Acqua et al., 2023). A potential explanation could be the difference in system usage behavior. In the exploratory analysis, we found low-creativity humans had made more requests for GLM suggestions, and of those higher numbers of ideas suggested, a higher share of ideas was adopted (with or without adjustments). Potentially, participants

compared their own ideas to the GLM suggestions. Given that GLMs can produce ideas comparable to many humans, it might not be surprising that humans with lower creative ability might view the GLM ideas more positively, adopt a higher share of the ideas, and request more additional ideas. In contrast, high-creativity individuals might be more critical of the GLM suggestions and adopt fewer of those ideas. Requesting new GLM ideas and reviewing them requires time. High-creativity participants need to find a good trade-off between coming up with own (potentially highly-creative) ideas and reviewing (potentially more average) GLM ideas. This might lead them to make fewer requests themselves. Besides such rational factors, emotional factors might be relevant. Prior studies found hints that humans who think of themselves as highly creative might be less willing to accept AI-support in a creative task (Memmert et al., 2023). Future research needs to explore this aspect further.

Understanding how working with the GLM affects humans (including potentially negative effects) is an important aspect for IS research (Adam et al., 2024). This might include more long-term effects of de-skilling due to an overuse of AI (Rafner et al., 2021), or a (short-term) reduction in effort when working with GLMs (Liu et al., 2023). In our study, we find GLM-supported humans to themselves produce fewer ideas regardless of creativity level, with an overall reduction of about 36.8%. Given that past research showed the most creative ideas to still come from humans, introducing a GLM might come at the cost of humans contributing less of their creative ideas, particularly so, as research shows that the most-creative ideas arise at later times during the brainstorming sessions (when the more obvious ideas are covered). Particularly problematic might be that the reduction of human fluency was significant for high-creativity participants. Thus, potentially, instead of contributing own original ideas, they might adopt mediocre GLM ideas. Furthermore, for professional work settings, originality, while a common measure for idea generation sessions, might not be the only outcome metric. Humans might be able to incorporate situation-specific information inaccessible to the GLM, leading to more tailored solutions, whereas working with the GLM might lead to more homogenous results (Dell'Acqua et al., 2023). Thus, the introduction of the GLM and the resultant reduction of human fluency might come at a cost even for low-creativity humans (e.g., if possessing situation-specific knowledge).

GLMs can produce new ideas at a rapid pace. Low-creativity participants made use of this GLM capability. This could hint at potentially undesired negative side effects like social loafing behavior, i.e., a reduction in effort. However, producing many ideas might also be considered positive, given that creative ideas might arise as recombinations and adjustments of prior ideas (Osborn, 1953), and typically occur later in the brainstorming session when the more obvious ideas have been expressed. Thus, potentially using the GLM to produce many ideas might not only help to quickly cover the obvious answers, but might lay the foundation to develop original ideas. In our study, we confirm earlier findings of a positive correlation between fluency and originality (Nijstad et al., 2010). However, for

positive, stimulating effects to materialize, humans need to pay attention to the suggested ideas (Dugosh et al., 2000).

ICSS can help humans in developing ideas. There are different types of ICSS that use different mechanisms to enhance brainstorming performance. Stimuli providers are a particularly effective type of ICSS (Althuizen & Reichel, 2016), that provide users with cues aiding them to generate ideas. Humans are known to run out of ideas during brainstorming sessions, or they might encounter fixation, i.e., focus on a narrow problem aspect. External stimuli might provide the necessary cues for the human to come up with new ideas, or overcome fixation by exploring different areas of the problem. Comparing the graphs of ideas produced within the GLM-supported and unsupported brainstorming sessions, one might suspect the GLM-support had such a positive effect on the humans, given the rate of ideas produced in the GLM-supported sessions decreased slower compared to the unsupported sessions (see Figure 5a). However, given the analysis on the effect of GLM support on human fluency, this might not be the case in our study, and instead might be caused by adopting GLM ideas (Figure 5b). Whether or not this is positive depends on the comparative quality of the ideas the participants could have produced instead of requesting and reviewing GLM ideas. The findings do, however, suggest that traditional stimuli-provider ICSS and GLM-based ICSS have different effects on the human, whereas the external information can increase human fluency in the former case, it can reduce fluency in the latter. Thus, depending on the goal—increasing individual human fluency or group level fluency—traditional ICSS or GLM-based ICSS might be more appropriate.

Our study ties into the existing literature on studying individual differences and how they affect benefiting from ICSS. As Wang and Nickerson (2017) pointed out, results as to whether the highly or the less creative benefit more from ICSS are inconclusive. In our study, while high-creativity humans benefit not in terms of fluency, both low- and high-creativity humans benefit in originality. Thus, whether or not the creativity level affects if humans benefited from GLM support, according to our results, depends on the goal of the brainstorming session. This particularly applies if fluency is only considered a means to arrive at a set of original ideas, or if having a high number of ideas to choose from is a goal in itself.

Stimuli-based ICSS usually used cues like words or (partial) ideas from dictionaries or social media. However, these inputs usually did not constitute a complete idea that can potentially be copied by the individual. In contrast, in our system, the GLM produces complete ideas comparable to human ideas, offering the option to edit or copy the ideas. We suggested, that our scenario potentially should be compared to human group brainstorming settings, where other group members can contribute ideas. While such ideas of others might have the a similar stimulating effect as stimuli providers – the effect is known as cognitive stimulation – such ideas might also invoke performance-reducing effects, such as social loafing (Liu et al., 2023), particularly when participants feel their effort is not needed. Thus, while GLM support was positively related to brainstorming performance across creativity

levels in our study, future research is required to develop a more nuanced understanding, e.g., regarding performance-reducing mechanisms.

16.5.2 Contribution to Theory

With our paper, we contribute to ICSS, group brainstorming, and human-GLM literature. GLMs are not only a change in the backend of ICSS, but due to being capable of producing complete, creative ideas enable new interaction opportunities for the human, from using the ideas as inspiration to editing and copying ideas. To study the dynamics in humans using such systems, we have developed a simple, GLM-based ICSS. Given the move towards being a colleague (Gabriel et al., 2016), we adopt a nuanced performance understanding differentiating individual and group level performance to dissect the effects.

We find GLM-based ICSS lead to improvements of group performance for both fluency and originality scores with medium to large effect sizes. On a surface level, the GLM seems to have reduced the natural fluency decline (Figure 5a), which seems to confirm earlier findings of the positive effect of stimuli providers (Althuizen & Reichel, 2016; Siangliulue, Chan, et al., 2015). However, we show that that unlike with traditional ICSS, *human* fluency did not increase but decrease, contrary to prior findings. Thus, applying prior knowledge from more traditional stimuli provider to GLM-based ICSS might be too simplistic, and a more nuanced perspective might be required. We discussed how known mechanisms from human groups might transfer onto human-AI groups offering potential explanations for the observed effects, and contributing to the discourse on human-AI group dynamics (Liu et al., 2023; Stieglitz et al., 2022).

Individual differences affect creativity (Mumford, 2012; Woodman et al., 1993), but the literature is inconclusive as to who benefits most from ICSS (Wang & Nickerson, 2017). With our study, we contribute to this discourse, showing that both the less- and the more creative can benefit, depending on the chosen metric. Analyzing the interaction data with the app, we have uncovered a difference in usage behavior between the low and high creativity participants, which might help to explain the results.

Additionally, we add to the growing body of literature on human-GLM (creative) work (Dell'Acqua et al., 2023). Several studies have explored different aspects of human-GLM brainstorming. More recently, focusing explicitly on GLMs, studies included comparing human and GLM brainstorming capability without interaction between human and GLM, i.e., in isolation (Haase & Hanel, 2023) and improving the creative output of GLMs on a technical level (Summers-Stay et al., 2023). Going beyond an isolated comparison, research was called for to investigate joint human-GLM brainstorming sessions (Di Fede et al., 2022), and on effects of working with GLMs on humans in general (Adam et al., 2024), and the effect of the task ability on performance when working with GLMs specifically (Liu et al., 2023). With our study, we contribute towards answering those calls for research. We confirm results of Dell'Acqua et al. (2023) in a classical brainstorming experiment setting using established brainstorming performance metrics. Similar to Dell'Acqua et al.

(2023), we find that creative ability does affect how people use and benefit from using GLMs.

With our study, we connect two streams of literature: ICSS research, which has a rich, decades-long body of literature, and, research on GLMs and creative knowledge work, which recently emerged and grows rapidly. An aspect both research streams seek to understand is how task ability affects usage behavior and task performance when being supported by software in creative work, which we have explore in our study.

Lastly, achieving a superior performance of the human-AI team compared to the individual is a key concern of human-AI collaboration research (Dellermann et al., 2019). With our study, we showcase a setting in which dyads of human and AI perform better than unaided humans, consistent with prior research (Memmert & Bittner, 2024). We discuss potentially negative side effects of using GLMs, relating it to the discussion in IS around the applicability of performance-increasing and performance-reducing group mechanisms known from purely human groups to human-AI groups (Liu et al., 2023; Siemon et al., 2015; Stieglitz et al., 2022).

16.5.3 Implications for Practice

Given our results, one might encourage the integration of a GLM capability into software used to support humans when generating ideas, as such a capability can improve both fluency and originality. However, users need to be made aware about the potential negative side-effects of using such systems, including a potential reduction in own ideas contributed. Users should reflect on their potential contribution considering their creative ability as well as their situation-specific (unique) knowledge and use the system accordingly. If possible, users should be supported in finding a good balance between contributing own ideas and curating the GLM's output. Given that the best ideas still originate from humans (Haase & Hanel, 2023), finding this balance might be particularly challenging for more creative humans, as they need to trade off contributing high-quality ideas themselves or reviewing (potentially more average) GLM ideas.

16.5.4 Limitations & Further Research

Our study has several limitations. To measure creative ability, we used the well-established approach of measuring divergent thinking ability (Olson et al., 2021). The measurement instrument DAT was selected, as it was shown to correlate with more established creativity measures (Olson et al., 2021), but can be completed in a short amount of time—beneficial also from a practical perspective if implemented into tools as a live assessment. While our participants might not be representative of the general population, our participants' measured DAT scores are consistent with the expected range of DAT scores reported by Olson et al. (2021). While we did not find a significant effect of measured creative ability on the dependent variables, a correlation analysis showed a significant, positive correlation between the DAT score and the originality score with a medium effect size (Pearson's

$r=0.330$, $p=.011$). A potential explanation might have been the limited sample size, which might have prevented us from detecting some significant differences, which is why we have reported the marginally significant results as well as the correlation between DAT score and originality score for completeness. Additionally, different creativity measures might have yielded different creative ability scores and, thus, affected analysis results. For the evaluation of brainstorming performance, we used two central brainstorming performance metrics: fluency and originality (Haase & Hanel, 2023). For rating originality, we adopted the approach of Siangliulue, Arnold, et al. (2015) in having participants select their best ideas and having an expert rate the ideas on a Likert scale. As a selection by participants could bias the results, we reduced subjectivity by providing the participants and the expert judge with the same, well-established definition of originality, which we adopted verbatim from the literature (Runco & Jaeger, 2012). We had only one judge rate the ideas, which might introduce subjectivity. Besides using a highly experienced expert not involved in the study and offering the same clear definition of originality for rating the ideas as was provided to the participants, we ensured the expert was blind-to-condition and the presentation order randomized. We undertook several measures to conceal if the ideas originated from the human or AI, e.g., by correcting grammatical errors or removing colloquial language. Additionally, we have investigated only one brainstorming question, one prompt template and one GLM (GPT-4). However, we have not adjusted the GLM to the specific brainstorming question, and studies show that GLMs can perform well in creative tasks across different topics (Haase & Hanel, 2023), that reasonable prompt variation has a limited effect on idea quality (Memmert et al., 2024), and that GLM ideas can improve group-level brainstorming performance (Memmert & Bittner, 2024). We have selected GPT-4 due to its good performance in creative idea generation (Haase & Hanel, 2023). However, we encourage future research to reproduce the results in different contexts with different GLMs to test result robustness. Lastly, we have not yet implemented design improvements on the various aspects of GLM support, thus, the effects of GLM support on performance might be considered a lower bound.

Future research might explore several sociotechnical design options to further enhance the GLM support, to combine the positive effects of traditional ICSS (increase human performance) and GLM-based ICSS (increase collective, human-GLM performance). This might include implementing GLM support along creativity theories, such as the dual pathway theory (Althuizen & Reichel, 2016; Nijstad et al., 2010), according to which humans can achieve more creative ideas, by moving along and switching between exploring the problem more broadly (flexibility pathway) and more deeply (persistence pathway). GLMs could, through appropriate prompt engineering, support humans on a specific pathway by providing only relevant ideas, and through appropriate timing of those ideas, trigger the human to switch pathways. Additionally, such systems might adapt to the creative ability of humans. GLMs can not only be used to generate ideas, but also for filtering (Summers-

Stay et al., 2023) and rating ideas (Haase & Hanel, 2023; Organisciak et al., 2023) automatically. Thus, the app could use a live pre-test similar to DAT used in our study, or rate the ideas of the participant as well as the GLM's ideas, prior to suggesting GLM ideas to the participant. Such a decision to suggest ideas should incorporate accumulated knowledge on brainstorming group mechanisms like cognitive stimulation, cognitive loafing, and social comparison (Liu et al., 2023; Pinsonneault et al., 1999; Stieglitz et al., 2022). This might include mechanisms to increase engagement with suggestions, which is known to be relevant to induce cognitive stimulation (Dugosh et al., 2000). In a follow-up analysis future research might explore how human ideas change in response to (or after) receiving outside stimulation from GLM ideas (e.g., using natural language processing).

16.6 Conclusion

With the increasing adoption of GLMs for creative work, understanding how individual differences such as creative ability affect how humans work with GLM-based systems and how this impacts creative outcomes becomes important. In our experiment, we observe that low-creativity humans seem to use the GLM-functionality of our app more intensely and to benefit across both brainstorming performance dimensions fluency and originality, whereas high-creativity humans only benefit regarding originality. However, while this might be a performance improvement, there seems to be a side effect of using a GLM: a reduction in ideas contributed by the human. We make suggestions as to how an appropriate sociotechnical design might help to realize the benefits of working with GLMs while reducing the unintended side-effects.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

16.7 References

- Adam, M., Bauer, K., Jussupow, E., Benlian, A., & Stein, M.-K. (2024). Generating Tomorrow's Me: How Collaborating with Generative AI Changes Humans. *Business & Information Systems Engineering*.
- Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of Management Information Systems*, 33(1), 11–44. <https://doi.org/10.1080/07421222.2016.1172439>
- Baruah, J., & Paulus, P. B. (2016). The Role of Time and Category Relatedness in Electronic Brainstorming. *Small Group Research*, 47(3), 333–342. <https://doi.org/10.1177/1046496416642296>
- Benbya, H., Strich, F., & Tamm, T. (2024). Navigating Generative Artificial Intelligence Promises and Perils for Knowledge and Creative Work. *Journal of the Association for Information Systems*, 25(1), 23–36. <https://doi.org/10.17705/1jais.00861>

- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Pers. Soc. Psychol.*, 53(3), 497–509. <https://doi.org/10.1037/0022-3514.53.3.497>
- Dugosh, K. L., Paulus, P. B. [P. B.], Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *J. Pers. Soc. Psychol.*, 79(5), 722–735. <https://doi.org/10.1037/0022-3514.79.5.722>
- Gabriel, A., Monticcolo, D., Camargo, M., & Bourgault, M. (2016). Creativity support systems: A systematic mapping study. *Thinking Skills and Creativity*, 21, 109–122. <https://doi.org/10.1016/j.tsc.2016.05.009>
- Gero, K. I. (2023). *AI and the Writer: How Language Models Support Creative Writers*. <https://doi.org/10.7916/wspj-fn77>
- Goss-Sampson, M. A. (2022). *Statistical Analysis in JASP 0.16.1*. <https://jasp-stats.org/wp-content/uploads/2022/04/Statistical-Analysis-in-JASP-A-Students-Guide-v16.pdf>
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706. <https://doi.org/10.1037/0022-3514.65.4.681>
- Kohn, N. W., & Smith, S. M. (2011). Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology*, 25(3), 359–371. <https://doi.org/10.1002/acp.1699>
- Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13(1), 13601. <https://doi.org/10.1038/s41598-023-40858-3>
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.*, 37(6), 822–832. <https://doi.org/10.1037/0022-3514.37.6.822>

- Liu, M., Ke, W., & Xu, D. J. (2023). Will Humans be Free-Riders? The Effects of Expectations for AI on Human-AI Team Performance. *PACIS 2023 Proceedings*.
- Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. <https://doi.org/10.1016/j.jbusres.2020.07.045>
- Memmert, L. (2024). Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects. In *Proceedings of the 32nd European Conference on Information (ECIS)*, Paphos, Cyprus. https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1
- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/107289>
- Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. <https://aisel.aisnet.org/icis2024/aiinbus/aiinbus/7/>
- Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429
- Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. *Journal of Information Systems Education*, 34(3), 279–292. <https://aisel.aisnet.org/jise/vol34/iss3/3>
- Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 589–612).
- Mumford, M. D. (Ed.). (2012). *Handbook of Organizational Creativity*. Elsevier. <https://doi.org/10.1016/C2009-0-01842-X>
- Nijstad, B. A., Dreu, C. K. W. de, Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, 21(1), 34–77. <https://doi.org/10.1080/10463281003765323>

- Nijstad, B. A., & Stroebe, W. (2006). How the group affects the mind: A cognitive model of idea generation in groups. *Personality and Social Psychology Review*, 10(3), 186–213. https://doi.org/10.1207/s15327957pspr1003_1
- Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences of the United States of America*, 118(25). <https://doi.org/10.1073/pnas.2022340118>
- Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49. <https://doi.org/10.1016/j.tsc.2023.101356>
- Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.
- Paulus, P. (2000). Groups, Teams, and Creativity: The Creative Potential of Idea-generating Groups. *Applied Psychology*, 49(2), 237–262. <https://doi.org/10.1111/1464-0597.00013>
- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., & Sherson, J. (2021). Deskillling, Upskilling, and Reskilling: a Case for Hybrid Intelligence. *Morals & Machines*, 1(2), 24–39. <https://doi.org/10.5771/2747-5174-2021-2-24>
- Rix, J. (2022). From tools to teammates: Conceptualizing humans' perception of machines as teammates with a systematic literature review. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.
- Runco, M. A. (2004). Creativity. *Annual Review of Psychology*, 55, 657–687. <https://doi.org/10.1146/annurev.psych.55.090902.141502>
- Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, 24(1), 92–96. <https://doi.org/10.1080/10400419.2012.650092>
- Siangliulue, P., Arnold, K. C., Gajos, K. Z., & Dow, S. P. (2015). Toward Collaborative Ideation at Scale. In *CSCW '15 Proceedings* (pp. 937–945). ACM. <https://doi.org/10.1145/2675133.2675239>
- Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In *C&C '15: Creativity and Cognition* (pp. 83–92). ACM. <https://doi.org/10.1145/2757226.2757230>
- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.

- Siemon, D., & Wank, F. (2021). Collaboration With AI-Based Teammates – Evaluation of the Social Loafing Effect. In *PACIS 2021 Proceedings*. <https://aisel.aisnet.org/pacis2021/146/>
- Steiger, T., Durani, K., Just, J., Hutter, K., & Eckhardt, A. (2024). From Concept to Creation: Artificial Intelligence in Innovation Teams. *PACIS 2024 Proceedings*. https://aisel.aisnet.org/pacis2024/track13_hcinteract/track13_hcinteract/16
- Stieglitz, S., Mirbabaie, M., Möllmann, N. R. J., & Rzycki, J. (2022). Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. *Information Systems Frontiers*, 24(3), 745–770. <https://doi.org/10.1007/s10796-021-10201-0>
- Summers-Stay, D., Voss, C. R., & Lukin, S. M. (2023). Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.
- Wang, K., & Nickerson, J. V. (2017). A literature review on individual creativity support systems. *Computers in Human Behavior*, 74, 139–151. <https://doi.org/10.1016/j.chb.2017.04.035>
- Williams, K. D., & Karau, S. J. (1991). Social loafing and social compensation: The effects of expectations of co-worker performance. *Journal of Personality and Social Psychology*, 61(4), 570–581. <https://doi.org/10.1037/0022-3514.61.4.570>
- Woodman, R. W., Sawyer, J. E., & Griffin, R. W. (1993). Toward a Theory of Organizational Creativity. *The Academy of Management Review*, 18(2), 293. <https://doi.org/10.2307/258761>
- Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2021). The effect of rank-ordering strategy on creative idea selection performance. *Eur. J. Soc. Psych.*, 51(2), 360–376. <https://doi.org/10.1002/ejsp.2743>

17 Publication 8: Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies

Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. *Journal of Information Systems Education*, 34(3), 279–292. <https://aisel.aisnet.org/jise/vol34/iss3/3>

Abstract

Design science research (DSR) is taught in university courses or used by students for their final theses. For successfully learning DSR, it is important to learn to apply it to (real-world) problems. However, students not only need to learn the new DSR paradigm (meta-level) but also need to develop an understanding of the problem domain (content-level). In this paper, we focus on content-level support (CLS), proposing an illustrative tool to aid students when learning to develop a conceptual design with DSR (e.g., for a prototype). Following the DSR paradigm, we deductively identify students' issues and use the scaffolding approach to develop design requirements (DRs) and design principles (DPs). To offer AI-generated scaffolding, we use the generative language model (GLM) 'GPT-3'. We evaluate our illustrative design through 13 expert interviews. Our results show that providing students with CLS is perceived to be helpful, but the interaction with the student needs to be designed carefully to circumvent unintended usage patterns. We contribute DPs and an illustrative instantiation thereof toward a DSR tool support ecosystem. More broadly, we contribute to the understanding of how humans can be supported by AI to solve problems, an important challenge in human-AI collaboration research.

Keywords: Problem-solving, Artificial intelligence, Scaffolding, Design science research, Generative language models, GPT 3

17.1 Introduction

Design Science Research (DSR) is taught to students in university courses (Hevner, 2021; Winter & vom Brocke), used as an educational approach to solve domain-specific problems (Thuan & Antunes, 2022), or used by students for their final thesis (Knauss, 2021). To effectively learn DSR, besides learning the theory, students need to *experience* DSR by applying it to (real-world) problems (Goldkuhl et al., 2017; Winter & vom Brocke), confronting them with both *meta-level* knowledge regarding the DSR paradigm and *content-level* knowledge regarding the problem domain (Figure 1). Learning a new paradigm while navigating a new problem domain can be challenging.

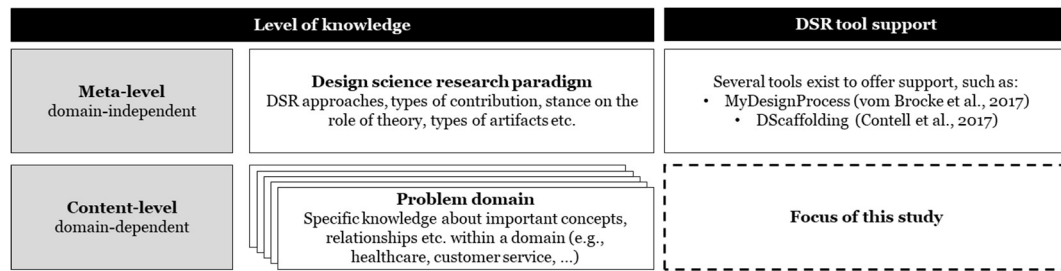


Figure 1. Focus of This Study With Respect to Levels of Knowledge

An approach from educational theory to support students in challenging learning situations is *scaffolding* (Wood et al., 1976). Scaffolding aims to offer support such as learning materials or cues to enable students to solve problems they otherwise could not solve. Tool-based scaffolding was proposed to address several learning challenges (Law et al., 2020). Existing tools for DSR support (e.g., Contell et al., 2017; Gau et al., 2022; Morana, Scheid, et al., 2018; vom Brocke et al., 2017) focus on problem-independent *meta-level* support, e.g., a canvas for organizing study contents or links to reading material. Offering CLS, however, might be particularly helpful to novices, as they might lack the required prior knowledge to grasp the problem and develop appropriate solutions. CLS could include suggestions concerning problem or solution aspects (e.g., underlying issues, DPs). However, preparing them for all topics a student might choose might not seem feasible.

With artificial intelligence (AI), however, we argue that CLS might be possible. In this article, we explore how it can help students to apply DSR, more specifically, in the aspect of developing a conceptual design. We answer the research question:

How can we use AI-generated, content-level scaffolding to support students in developing their conceptual design with DSR?

To answer this question we develop an illustrative prototype following DSR. It provides scaffolding to address issues of students when solving problems, offering content-level suggestions through a GLM. We evaluate our illustrative prototype through 13 expert interviews. Overall, interviewees liked the idea of CLS to aid students, highlighting the opportunity of receiving new ideas to challenge and improve the developed concept. We discuss required research on undesired behavior when using such generative AI. We evaluate the AI suggestion quality in a different, quantitative study, therefore it is out-of-scope for this paper.

With DPs and a situated instantiation we contribute evaluated design knowledge towards a DSR tool support ecosystem as suggested by Morana, vom Brocke, et al. (2018) and to the understanding of AI-based scaffolding opportunities in higher education, a trend suggested by Doo et al. (2020). More broadly, due to our deductive approach we contribute towards a better understanding of how humans and AI systems can collaborate to solve problems, an important challenge for human-AI collaboration research (Dellermann et al., 2019). For this paper, we mostly follow the publication scheme for DSR papers by Gregor and Hevner (2013).

17.2 Background

We explain *scaffolding* as conceptual and GLMs as technical foundation, with GLMs enabling the dynamically generated, context-specific scaffolding. We then relate our proposed tool to existing DSR support tools.

17.2.1 Scaffolding

Scaffolding is the "process that enables a child or novice to solve a problem, carry out a task or achieve a goal which would be beyond his unassisted efforts" (Wood et al., 1976, p. 90). This not only allows novices to successfully complete a task by concentrating on the aspects they can handle but also enables them to build "task competence [...] at a pace that would far outstrip [their] unassisted efforts" (Wood et al., 1976, p. 90). Given our goal to support students in *learning to solve a problem* using DSR, supporting both *problem-solving* and *learning* through scaffolding seems appropriate. Wood et al. (1976, p. 98) explain several scaffolding functions such as fostering the learner's interest, maintaining their focus and motivation, reducing the task's "degrees of freedom" and "marking [the task's] critical features" to challenge the learner, and, demonstration, which "may involve completion or even explication of a solution already partially executed by the tutee himself". M. C. Kim and Hannafin (2011, p. 408) identified different purposes such as procedural- (operational aspects), conceptual- (knowledge gaps), and strategic scaffolds (alternative approaches), which we refer to during design and discussion.

Scaffolding was discussed in real-life, "authentic learning tasks", as, given the "complexity of those tasks, learning may be hampered by the limited processing capacity of the human mind" (van Merriënboer et al., 2003, p. 5). As DSR is used to solve such complex, real-life problems (Hevner et al., 2010), scaffolding might thus be particularly suited. We, therefore, employ scaffolding for developing requirements of and principles for our solution. Seeking to explore the helpfulness of AI-generated scaffolding, we focus on offering (instead of fading) support.

17.2.2 Generative Language Model (GLM)

GLMs are a specific type of AI trained to predict an input text's (likely) continuation, producing a corresponding output. Figure 2 shows three examples of responses of the GLM 'GPT-3' (Brown et al., 2020) to our inputs. Importantly, the model does not need training (data) with these inputs in mind.

Human prompt	a)	INPUT	What is the capital of the USA?
		OUTPUT	Washington, D.C.
		AI response	
	b)	INPUT	Man -> King Woman ->
		OUTPUT	Queen
		AI response	
	c)	INPUT	A conversational agent is a dialog system that uses natural language. In order for conversational agents to not be perceived as uncanny (uncanny valley theory) they should satisfy the following design principles:
		OUTPUT	1. Be human-like: the agent should use natural language and be able to hold a conversation. 2. Be responsive: the agent should be able to understand and respond to the user in a timely manner. 3. Be adaptive: the agent should be able to adapt its behavior to the user's preferences. 4. Be error-tolerant: the agent should be able to handle errors gracefully.
		AI response	

Figure 2. Authors' Exemplary Input Prompts and Outputs Generated by GPT-3

As students might select any topic for their DSR study, building a repository of training data in advance (e.g., all possible issues or DRs for any potential topic) might not be feasible. Because of its flexibility to accept arbitrary input and produce corresponding output across domains without requiring problem-specific training data, and its capability to inspire humans (Lee et al., 2022; Memmert & Bittner, 2022) and “accurately reflect important aspects of human conceptual knowledge” (Hansen & Hebart, 2022, p. 6), we selected GLMs for realizing AI-generated scaffolding. However, GLMs can cause challenges, e.g., the outputs might lack truthfulness (Lin et al., 2022) while seeming plausible, or worse performance when insufficient contextual information is provided (Floridi & Chiriatti, 2020) or when inexperienced users interact with such systems (Jiang et al., 2022). We reference these challenges during design.

17.2.3 DSR Tool Support

DSR tool support was discussed within the IS community (e.g., at DESRIST 2017) and requirements were gathered accompanied by a call for a DSR tool support ecosystem (Morana, vom Brocke, et al., 2018). A backward and forward search on this call and a search for "Design science tool", "Design science research tool", or "DSR tool" surfaced articles proposing design knowledge and tools such as "DScaffolding" (Contell et al., 2017), a mind-mapping-based tool targeted at novices to support learning DSR while conducting a study, or "MyDesignProcess.com" (Morana, Scheid, et al., 2018; vom Brocke et al., 2017), a canvas-based tool to collaboratively plan, work on, document, and communicate DSR studies. Extending the latter, even a conversational agent was proposed to reduce the documentation effort (Gau et al., 2022).

The tools mentioned above focus on supporting on generic, problem-independent meta-level, not on content-level (Figure 1). Assuming one develops a study concept for "*facilitating an organization's innovation process*", meta-level support might include, e.g., listing the three most common DSR approaches (problem-independent). Conversely, CLS could include suggestions for potential issues, e.g., "*ideas of stakeholders often lack the required level of detail to be helpful*" (problem-dependent). In this study, we seek to offer the latter, complementing and thereby expanding existing DSR tool support, striving towards an ecosystem (Morana, vom Brocke, et al., 2018).

17.3 Method

We follow the DSR approach by Kuechler and Vaishnavi (2012) to develop an (illustrative) prototype to aid novices in learning to apply DSR (Figure 3). To create *awareness for the*

problem, we conceptualize DSR as a problem-solving approach. Consequently, we can deductively use (the more general) problem-solving literature to identify issues arising when humans, particularly novices, solve problems. Based on these issues, we develop literature-guided *suggestions*. More specifically, we derive DRs and formulate DPs (following Meth et al., 2015) based on scaffolding literature, leveraging learner-focused frameworks for "scaffolding complex learning" and scaffolding "ill-structured problem solving" (Law et al., 2020; Reiser, 2004). We *develop* an illustrative prototype based on a GLM, as GLMs can create context-dependent text enabling AI-generated scaffolding.

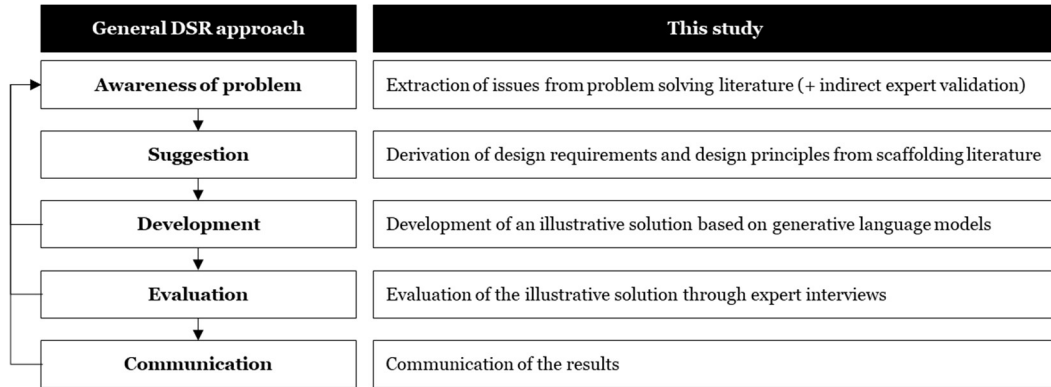


Figure 3. Our DSR Approach Based on Kuechler and Vaishnavi (2012)

Our evaluation aimed to understand whether our design was assumed to be helpful to novices for developing concepts. We performed semi-structured expert interviews along an interview guideline (Meuser & Nagel, 2009), covering aspects like the overarching structure and CLS (see appendix). After we explained the functionality and planned interaction along prototype visualizations, interviewees reflected on their experience with novices and how such a tool might be beneficial. On the transcripts, we performed a qualitative content analysis using "Content Structuring/Theme Analysis", a mixed procedure using both deductive and inductive coding (Mayring, 2014, p. 104). While not evaluated formally, during the interviews, the experts mentioned most of the issues we had identified from literature and used for tool development.

With regard to the DSR evaluation framework of Venable et al. (2016) our evaluation can be considered between ex-ante and ex-post as a conceptual prototype with an illustrative design for the user interface and the approach for producing AI suggestions (suggestion quality evaluated in a different study (Memmert et al., 2023)) was already developed, but the system is not yet technically implemented e.g., generating AI suggestions during the study required manual population of the template strings. It can be considered formative as we seek to "produce empirically based interpretations that provide a basis for successful action in improving the characteristics or performance of the evaluand" (Venable et al., 2016, p. 78), particularly given the idea's novelty and the unclear impact within a learning situation. We chose an artificial evaluation asking experts about the tool's potential impact, as experts are experienced with novices' challenges in general and might provide more fruitful feedback than students who might have only conducted one (partial) DSR study.

17.4 GLM-based DSR Support Tool

To learn DSR effectively, students need to solve a problem practically (Goldkuhl et al., 2017; Winter & vom Brocke). We highlight issues that might arise and offer suggestions to support the concept development aspect of a DSR project.

17.4.1 Awareness of the Problem

DSR is an approach to solving complex, ill-structured problems (Hevner et al., 2010; vom Brocke et al., 2020). Solving such problems can be difficult for novices (Fischer et al., 2012) and even more challenging while learning a new paradigm (i.e., DSR). Several issues may arise.

A **lack of prior domain knowledge (I1)** can make it difficult to identify the relevant concepts to organize the path from problem to solution space (Fischer et al., 2012). An insufficient discovery of important issues, requirements, principles, or theories can result in an incomplete DSR concept. Due to the lack of prior domain knowledge and the problem's open-endedness, students might require additional information. Typically, teachers prepare material upfront so students can focus on learning instead of gathering information. Students are appreciative if they can select a DSR study topic themselves (Thuan & Antunes, 2022) and are encouraged to do so to increase engagement (Winter & vom Brocke). Due to this flexibility, however, it might not be feasible to prepare CLS upfront for every topic a student might select due to time constraints or because teachers might not have expert knowledge. Thus, there is a **lack of content-level support (I2)**.

Consequently, students must gather information themselves besides developing the concept, resulting in frequent **task switching (I3)**. When students have to switch between the task and searching for additional information, the split-attention effect (Sweller & Candler, 1992) can occur, which can be cognitively demanding and affect the learning process, especially "if the subject matter is complex" (Grévisse et al., 2019, p. 3). Additionally, students might **not be very effective in searching relevant information (I4)** because, as novices, they might not know for which information to search or lack the required search terms (Grévisse et al., 2019). The search might also be ineffective, as students might **lack methodological training (I5)** in the information-gathering methodologies employed in DSR such as literature search (Karlsson et al., 2012). Due to their lack of prior knowledge, students might find it challenging to navigate the domain, choosing an **inappropriate level of abstraction (I6)** for their analysis. They might focus on salient aspects or superficial details while neglecting other, more important features (Law et al., 2020; Reiser, 2004). Lastly, while writing a final thesis or taking a university DSR course, students might face a **lack of time (I7)** to complete an entire study with literature analysis and empirical data gathering, a potential constraint highlighted by Cater-Steel et al. (2019) even for Ph.D. students given that "a significant DSR program typically encompasses many researchers over several years" (Gregor & Hevner, 2013, p. 339).

After the tool development, during the expert interview sessions for the tool evaluation, the general problem of increased complexity for learners due to learning both a new paradigm and domain was mentioned. While not systematically evaluated, the experts mentioned all but one (I3) of the issues. We assume I3 was not mentioned due to its operative nature. The split-attention-effect, however, is well documented in cognitive load literature (Sweller & Candler, 1992). Experts highlighted that issues vary across students and can (partially) be addressed through the supervisor.

17.4.2 Suggestion

Based on the identified, literature-guided issues and the scaffolding approach, we derive DRs for the solution and formulate DPs to address them (Figure 4).

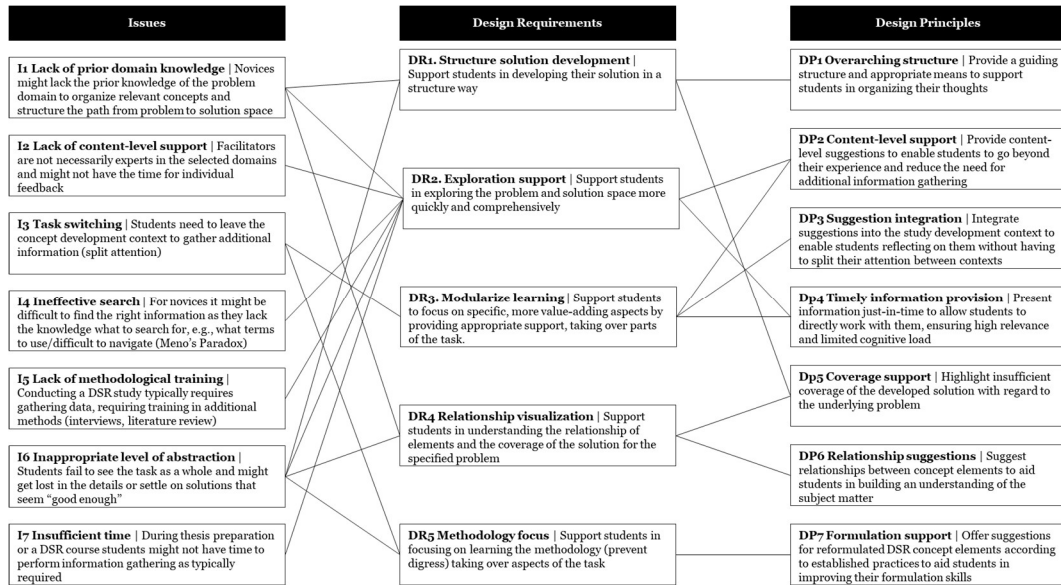


Figure 4. DSR Study Concept

Students lacking prior domain knowledge and only just learning DSR might find it challenging to organize their thoughts on the right level of abstraction (I1,I6). A common means is to provide a guiding structure like a process or template to highlight the key elements (Law et al., 2020). This might be considered a *hard* scaffold, which refers to "static supports that can be anticipated and planned in advance based on typical student difficulties with a task" (Saye & Brush, 2002, p. 81). For ill-structured problems, this might aid students, preventing an oversimplified problem and solution representation (Law et al., 2020). Therefore, the tool should support students in **structured solution development (DR1)**. Thus, we propose:

DP1 Overarching structure. The tool should provide a guiding structure and appropriate means to support students in organizing their thoughts.

To develop sound solutions, the problem and solution space – reflected in our canvas through pre-defined DSR-specific concept elements – needs to be explored appropriately, which might be difficult for students (I1,I4,I6,I7). *Conceptual scaffolds* can support stu-

dents in identifying knowledge gaps, and *strategic scaffolds* in considering alternative perspectives on their “preliminary or tentative solutions” (M. C. Kim & Hannafin, 2011, p. 408), preventing adopting “oversimplified misconceptions” (M. C. Kim & Hannafin, 2011, p. 412). Typically, hints, prompts, or strategic tips can be provided by teachers to challenge students and aid them moving forward (M. C. Kim & Hannafin, 2011; Reiser, 2004; van Merriënboer et al., 2003). These might be considered soft scaffolds, as they are “dynamic and situational” (Saye & Brush, 2002, p. 82). However, instructors might not always be able to provide problem-specific support given the wide range of potential topics and limited resources (I2). Therefore, students should receive **exploration support (DR2)**. Additionally, to prevent students from having to switch between developing the concept and gathering additional information (I3), students should be supported to focus on specific aspects of the task through **modularizing learning (DR3)**. This “reduction in degrees of freedom” (Wood et al., 1976, p. 98) by taking over aspects of the overall task should help to maintain focus. We therefore propose:

DP2 Content-level support. The tool should provide content-level suggestions to enable students to go beyond their experience and reduce the need for additional information gathering.

Content-level suggestions are additional information to be considered, which might increase cognitive load. To reduce cognitive load, AI-generated scaffolds are integrated directly into the tool, as otherwise, the split-attention effect (Sweller & Candler, 1992) might occur. We therefore propose:

DP3 Suggestion integration. The tool should integrate suggestions into the study development context to enable students reflecting on them without having to split their attention between contexts.

As additional information could potentially increase cognitive load, van Merriënboer et al. (2003, p. 9) recommend presenting information of lower complexity “precisely when learners need it during their work on the learning tasks”, as the risk of cognitive overload is smaller. We therefore propose:

DP4 Timely information provision. The tool should present information just-in-time to allow students to directly work with them, ensuring high relevance and limited cognitive load.

Due to the students' potential lack of domain knowledge and incorrect level of abstraction, students might struggle to appropriately structure the path from problem to solution space, e.g., focusing only on certain aspects of the problem or solution (I6). However, appropriate scaffolding should enable students to think about “disciplinary relations” (Quintana et al., 2004, p. 347). Therefore, students should be supported in appropriately covering the solution for the specified problem through **relationship visualization (DR4)**. We therefore propose:

DP5 Coverage support. The tool should highlight insufficient coverage of the developed solution regarding the underlying problem.

Whereas experts can organize knowledge effectively within mental schemas, understanding the relationships between concepts can be difficult for novices (Fischer et al., 2012). Conceptual scaffolding can help "students consider tasks from different angles through the reorganization and connection of evidence" (N. J. Kim et al., 2018, p. 402). To increase transparency in solution development, relationship visualization should actively support students in linking conceptual elements by making suggestions for potential links, challenging them to further explore and structure the problem and solution space. We therefore propose:

DP6 Relationship suggestions. The tool should suggest relationships between concept elements to aid students in building an understanding of the subject matter.

A common scaffolding technique is offering worked examples (van Merriënboer et al., 2003) or expert modeling (N. J. Kim et al., 2018), where an instructor performs the task correctly for demonstration. To support the students in learning DSR standards, such as correctly formulating DRs or DPs, meta-level support might link to reference literature. Building on this idea, students should be supported in applying established ways of phrasing concept elements to maintain **methodological focus (DR5)** and prevent digressing. Thus, we propose:

DP7 Formulation support. The tool should offer suggestions for (re-)phrasing DSR concept elements according to established practices to aid students in improving their formulation skills.

17.4.3 Development

To implement the DPs, we have developed an illustrative design (Figure 5). The overarching structure (DP1) is manifested through concept elements (issues, DRs, DPs) in a template-like manner (hard scaffold). Concept elements can be linked to show the path from problem to solution space. The idea of problem and solution space was discussed in DSR literature (Thuan et al., 2019; vom Brocke et al., 2020), and so was creating relationships between problem and solution space elements (Venable, 2006). Our visualization is adapted from published DSR studies (e.g., Gnewuch et al., 2017; Meth et al., 2015). For comprehensiveness, we added elements typically included in the studies' manuscript (situation, general problem, artifact class). Besides providing humans with a quick overview, these capture the contextual information enabling the tool's core proposition: content-level support (DP2).

GLM-based DSR Support Tool			
Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes			
Situation Organizations strive to leverage external expertise through digital idea platforms to improve innovativeness	General Problem Curating the generated ideas by external stakeholders requires high manual effort hindering organizations to exploit the full potential of open innovation initiatives	Artifact Class Conversational Agent	
Issues High human effort required to organize and evaluate ideas Well-elaborated, familiar ideas might be selected instead of innovative ideas Submitted ideas lack the required level of detail and richness Add your issue here...	Design Requirements Conversational agents needs to show human-like and reciprocal conversational behavior that is adjusted to the specific audience. Conversational agents should make approving and motivation statements during the process. Conversational agents should show politeness and act friendly. Add your design requirement here...	Design Principles Provide a conversational agent with the capacity to initiate, construct and lead a conversation in a productivity-oriented and pleasant manner by actively posing questions to reach the facilitation goal and considering strategies to prevent deviations to other topics. Reformulated according to Chandra et al. (2015) – changes in <i>italic font</i> Provide a conversational agent with the capacity to initiate, construct, and lead a conversation in a productivity-oriented and pleasant manner by actively posing questions to reach the facilitation goal and considering strategies to prevent deviations to other topics, <i>in order for users to engage in a productive conversation, given that the user wants to maintain a conversation about a specific topic.</i> Recommended structure: Provide the system with [material property—in terms of form and function] in order for users to [activity of user/group of users—in terms of action], given that [boundary conditions—user group's characteristics or implementation settings]. Provide the conversational agent with the capacity to actively react to and animate the user during the facilitation process to reach elaborated outcomes by offering support on demand with explanations about the process steps or answering topic-related vocabulary questions in a goal-oriented way. Provide the conversational agent with socio-emotional cues for a personalized interaction with humans by demonstrating appropriate, friendly, polite and human-like characteristics in the communication by utilizing informal language and delivering motivating statements to reach higher acceptance in user's perception as facilitator. Add your design principle here...	Theories Social Response Theory (Nass & Moon, 2000) Facilitation Framework (Bostrom et al., 1993) Social Presence Theory (Daft & Lengel, 1986; Gefen & Straub, 1997; Short et al., 1976) Add your theory here...
AI Suggestions: Issues  The evaluation process is often biased and subjective The evaluation process is time-consuming The ideas generated might not be useful for the organization	AI Suggestions: Design Requirement  Conversational agents should elicit ideas by asking open-ended questions. The conversational agent should be able to guide the user through the process by providing structure and support. The conversational agent should be able to evaluate the ideas submitted by the user and provide feedback.	AI Suggestions: Theories  Social Information Processing Theory (SIP) (Walther, 1996) Computer-Supported Cooperative Work (CSCW) (Bachman & Whitaker, 1993) Task-Technology Fit Theory (Goodhue & Thompson, 1995)	

Figure 5. Illustrative Design of our GLM-based DSR Support Tool (Staged with Exemplary Contents Cited or Paraphrased From Poser et al., 2022)

Content-level support (DP2) is realized through context-specific suggestions. Unlike the pre-defined overarching structure (DP1, hard scaffold), these are generated dynamically during concept development, based on the current concepts' contents (soft/dynamic scaffold). To achieve such functionality, we employ a GLM. However, learners might not be familiar with GLMs, which can lead to worse output quality (Jiang et al., 2022). Thus, we do not let learners interact with the GLM directly but automate the request in the backend. The information the human entered in the tool's upper part is plugged into pre-defined templates and provided to the GLM, which generates suggestions displayed in the bottom part to be reviewed (and potentially integrated into the concept) by the human (see Figure 6). As insufficient contextual information might reduce the GLMs performance (Floridi & Chiriatti, 2020), we use the study setting information captured in DSR-specific concept elements of our canvas (situation, general problem) as additional input when generating suggestions. All suggestions for all concept elements in Figure 5 are AI-generated with this approach. For this study, the process was performed manually.

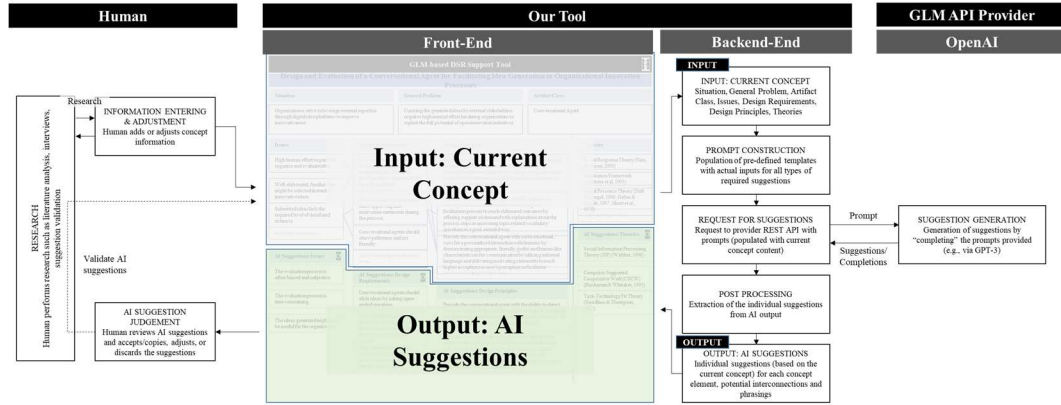


Figure 6. Schematic Flow of Information: Current Concept State (Upper Part) is Filled into Pre-Defined Templates Resulting in Prompts Fed to the GLM (Provider API) to Generate Suggestions. Cleaned Suggestions are Reviewed by the Student

Modularization (DP3) is achieved by integrating AI support directly into the tool as condensed, itemized suggestions, displayed directly below the current concept. Reducing the need to leave the tool to search for additional information, students can thereby include the suggestions directly into the concept as they wish (copy, adjust, or ignore). We offer timely suggestions (DP4) by allowing the student to refresh the suggestions themselves, i.e., just-in-time when ready to review them. The tool offers coverage support by highlighting concept elements without connections (DP5). While the student connects the concept elements to build a coherent concept, the tool suggests potential links (DP6) based on the concept elements' texts' (CLS), challenging the student to reflect. Lastly, the tool provides formulation support (DP7) for concept elements. The student can click on concept elements to see suggestions for re-phrasing them following established practices. Receiving the re-phrasing task, the original human input, and the target structure (e.g., DP structure of Chandra et al., 2015), the GLM generates the suggestions (see Figure 7).

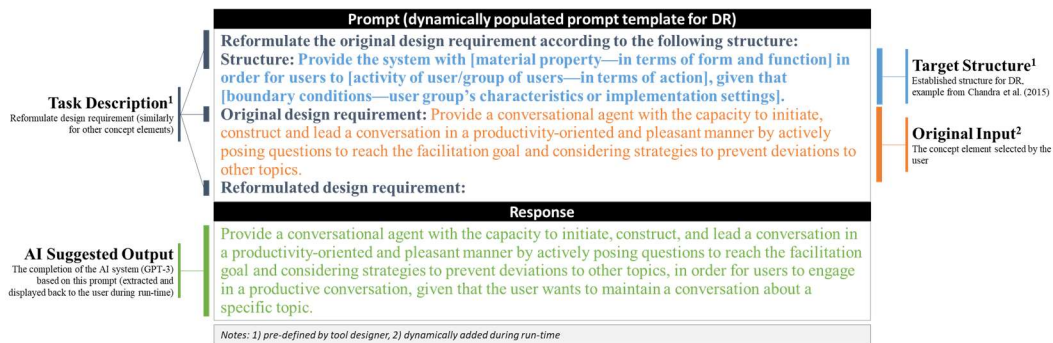


Figure 7. Automatically Assembled Prompt with a Pre-Defined Task Description and Target Structure, and Dynamically Added Student Input, Producing AI Output

To effectively use the tool, students need an introduction to DSR, developing concepts using interlinked concept elements, and the tool's functionality. This includes highlighting the GLM's limitations, particularly the risk of plausible-sounding but incorrect information (Lin et al., 2022) and the necessity to treat them as hypotheses to be reflected on. For concept development, the student adds content (e.g., situation, general problem, issues) onto the empty canvas based on their initial research (e.g., literature analysis, data gathering).

This serves as input for the GLM, which generates suggestions, challenging the student to develop the concept further. Adding new content, the student iteratively improves the concept, requesting suggestions when desired.

While the evaluation of suggestions was out of scope for this study (part of a different, quantitative study), we staged the tool with exemplary contents from a published DSR study to help the experts understand the tool idea during the interviews. However, the tool is not adjusted to this specific study. We intentionally did not formulate design features given the high uncertainty of the utility of this novel, AI-based scaffolding approach but instead chose an illustrative design to gather feedback.

17.5 Evaluation

Seeking to determine whether the design is assumed to be helpful to novices, we conducted 13 interviews with Ph.D. students, postdocs, and professors (see Table 1) working at eight universities across four countries. All interviewees are experienced in using DSR and working with novices supervising students' theses or teaching DSR courses. Quotes from interviews held in German were translated by the authors.

Table 1. Interviewee Details

#	Position	DSR Experience			Gender	Interview length [mins]
		Used in own publications	Supervised students	Taught courses DS(R)		
E1	PhD candidate	X	X		M	~41
E2	PhD candidate	X	X		M	~55
E3	PhD candidate	X	X		M	~51
E4	PhD candidate	X	X		F	~39
E5	Postdoc	X		X	M	~44
E6	PhD candidate	X	X		F	~43
E7	PhD candidate	X	X		M	~32
E8	Professor	X		X	M	~36
E9	Professor	X	X		M	~28
E10	PhD candidate	X	X		M	~45
E11	Postdoc	X	X		M	~37
E12	Professor	X		X	M	~53
E13	Postdoc	X	X		F	~50

Overall, the tool was perceived positively. Interviewees highlighted its potential to both support students in improving the DSR concept's quality and to better learn the DSR paradigm. Interviewees, though, raised concerns regarding potential undesirable usage patterns. We discuss the results along the DPs, in alignment with the thought process during the interviews, with a focus on the first two DPs.

Having an **overarching structure (DP1)** was perceived positively – “these are actually exactly the things that are important” (E9) – by many interviewees (e.g., E2,E3,E9), particularly to support novices (e.g., E2,E5,E9). It was seen as "guidance" or "process" offering "good structure" (E3,E9) that could increase the concept development speed (E4), encourage "justifying" design decisions (E12), and encourage starting with issues instead of the "typical pattern" of 'solution first' (E5). The large amount of DSR literature can overwhelm novices (E7). A tool could offer "better support" than sending exemplary DSR studies to students (E4), explicating the relevant questions to be answered (E9) or fields to be filled (E11). Being confronted with learning both about the problem domain and the paradigm can be overwhelming (E7). The tool could allow students to focus on the actual problem first while being led through the process, and only once a sufficient understanding is reached, use specific methods to dive deeper into the topic (E7). A fixed structure might be

more helpful for novices (e.g., E1,E2,E11), with one expert stating the structure to be “super helpful, especially for beginners, because I think it makes it easier for them to get started” (E2), as with more experience situation-appropriate adaptations are made (E5).

Many interviewees discussed the aspect of an overarching structure along our specific structure. Some interviewees recognized the flow from issues to DPs as familiar and helpful (e.g., E4,E8,E10) or as “essential part of DSR” (E2). However, several interviewees stressed that this exemplifies only one DSR approach and that there are other ways of conducting DSR and codifying design knowledge (e.g., E4,E11,E12). Our more practical approach is perceived to be in contrast to more theory-driven approaches, e.g., by Gregor (2006) (E12). Some critique was more fundamental, with one interviewee mentioning an unclear connection to current conversations around DSR (E12). Interviewees made several suggestions, including additional elements such as *stakeholders*, *environment*, or *organizational setting*, and *design features*, *design decisions*, or *design implementation* for better problem and solution space exploration, respectively (e.g., E1,E6,13).

Some interviewees explained that students sometimes lack required domain knowledge (I1; E2,E3) as topics are innovative and at most partially mentioned in teaching (E3). Supervisors usually provide suggestions also on content-level, e.g., constructs, literature, or keywords for literature search (e.g., E7,E9,E10). Thus, most interviewees liked having complementary **content-level support (DP2)**, e.g., stating “I find this enrichment of content exciting, and it is certainly helpful” (E2) or “the interesting thing in my opinion is that you have these suggestions, so ideas that are qualitative, as brainstorming ideas or as idea givers, which I find very very exciting” (E9). They saw AI suggestions as a source of inspiration, creative impulses, or new ideas (E1,E2,E4), expanding the concept but requiring validation (E2). Interviewees explained that some students lack methodological training (E4) for conducting interviews or literature search (I5;E1). Combined with insufficient domain knowledge, this can lead to ineffective literature search (I4;E2). Interviewees proposed suggestions could be “thought-provoking” (E10) or “pointers” (E2,E9) – “I think it's very good as a pointer in the sense of ‘this could be relevant’” (E9) – for a more targeted, explorative literature search or data gathering (e.g., E2,E6). Suggestions could thus potentially speed up the search (I7;E2) or even change the way of working towards “rapid prototyping” (E13).

Suggestions were seen as opportunities to engage in reflection (E2) on ideas and to prevent students from “marrying their first idea” (I6;E5). Particularly when students work alone, the tool could act as a “sounding board” (E5), encouraging an iterative approach, with one expert summarizing this thought as follows: “It also somehow fits in with DSR, because it has something iterative about it, maybe you only have 2 or 3 issues at first and maybe you have the first design requirements and you may or may not already have design principles; and then you get new theories suggested, then you would somehow engage with the literature again and then you could continue to work with your tool and add to it again. Then you would have work steps with the tool and without the tool. I think that's quite cool.”

(E10), with humans adding their initial ideas, receiving suggestions, conducting additional analysis, and refining their concept (E10). Supervisors sometimes do not have the time or topic knowledge (I2) to provide recommendations or feedback on a detailed level (E6), particularly when supervising a broad spectrum of topics beyond their core research (E6).

Interviewees saw potential for using the tool across knowledge levels, from novices (e.g., E1,E2) over Ph.D. students (e.g., E10) up to experienced DSR researchers (e.g., E1,E5), with varying usefulness assessments regarding concept elements, with one expert stating that if “you get feedback on [what you have written], then that is certainly also beneficial for getting better results. That doesn't have to be the case only for novices. I would also say that if I were to somehow muddle along on my own, I wouldn't mind being encouraged to not be satisfied with what I wrote the first time around, and that's how professional designers work too.” (E5). Using such a tool, however, requires understanding the underlying concepts, like differences between concept elements (e.g., E5,E8). Additionally, working with the tool requires discipline. Most interviewees raised the concern of novices accepting suggestions without reflection (e.g., E8,E13) using a “trial-and-error” approach (E4), refreshing suggestions until finding something plausible and superficially searching for matching sources, essentially “retrofitting” their solution (E8), with one expert raising the point that “depending on how good the tool will be later on, the question is whether the students can just use it, enter an example and generate their theses in the end” (E6).

Interviewees recommended preventing such unintended usage through design, e.g., by experimenting with timing (E3; see DP4), adding a reflection phase (E6,E13), marking AI-generated concept elements (E3), or adding a conversational agent encouraging engagement via a “Socratic dialog” (E8). Unintended usage was by far the most frequently raised concern. Others included, e.g., students' potential inability to provide sensible input (E8) or students not learning to work creatively and independently (E6).

Some interviewees discussed that the **integration of the suggestions (DP3)** makes knowledge more accessible than, e.g., having papers recommended by the supervisor, which can be overwhelming (E7). Some interviewees liked the “business model canvas”-like (E9) integrated one-pager (E1) while others found it overwhelming, particularly for extensive studies (E1,E6) and suggested incrementally displaying the steps (E9) in a “user flow” (E6).

Regarding the **timely information provision (DP4)** – in line with comments on DP2 – interviewees explained that providing suggestions all at once could lead to uncritical copying (E4). One interviewee, therefore, suggested investigating information provision timing and introducing a phase-based approach (E3).

Coverage support (DP5) and **relationship suggestions (DP6)** were presented together during the demonstration as both concern concept element linking. Coverage support was perceived as helpful (e.g., E1,E4,E5), as it can act as a “trigger” to justify design deci-

sions (E5) and thereby foster learning the DSR mindset (E1). Interviewees found relationship suggestions (DP6) helpful, particularly in a "narrowing down mode" (E11) with many concept elements, with one expert stating "sometimes you have so many design requirements that you do not have an overview of everything that fits together with the content of the design principles" (E10). Relationship suggestions were seen as help to identify argumentation gaps (I6) and triggers for reflection (E6), potentially sharpening arguments (E2,E6) and the overall concept (e.g., E2,E6,E7). A supervisor might not discuss on this micromanagement level (E5), but the tool could do so, improving understanding.

Lastly, concerning **formulation support (DP7)**, interviewees explained that formulating concept elements such as DPs is difficult for students (e.g., E2,E3,E4), e.g., sometimes DPs are too generic or specific (I6;E1). They explained that formulation suggestions for the student-entered content – additionally to the generic concept element structure – were beneficial (E7,E10) for understanding concepts like DRs or DPs, and thereby, DSR aspects (E11). One expert explained that even with formulation references and as an expert, one still sometimes "has difficulties formulating and squeezing it in [the structure]. That's why I think this feature alone is something that is very helpful" (E4). Some concerns mentioned previously were reflected here too (e.g., accepting unfit suggestions).

17.6 Discussion

17.6.1 Meta-level Support

Most interviewees agreed that an overarching structure ("DSR canvas") – in general – could be helpful, particularly for novices to understand the idea of DSR. Offering a structure is based on the concept of *hard scaffolding* (Law et al., 2020). It can be considered a procedural scaffold (M. C. Kim & Hannafin, 2011), reducing complexity and leaving the novice to focus on the concept development. However, there is dissent about how such a DSR structure should look like. Several experts commented positively, while others suggested additional elements or questioned the DP-based approach altogether. One interviewee explained that there are different "churches of DSR" (E12) and that developing a canvas is like opening a "can of worms" (E12). There are many views on DSR, and our tool adopts one view (adapted from several published DSR papers). We do not seek to make a methodological contribution to the DSR paradigm, i.e., we do not claim this structure to be a good or the best way to support novices. Given the community discussion, developing a DSR canvas is way beyond our paper's scope. However, we show that most experts agree on the need for scaffolding until students reach a certain DSR proficiency, allowing the supervisor to "offer feedback at a higher level instead of technical aspects" (E5).

17.6.2 Content-level Support

We use the canvas-like structure as a vehicle to offer content-level suggestions, which can be considered soft or dynamic scaffolding, a part of holistic scaffolding as proposed by

Law et al. (2020). Inspired by generative AI supporting humans in open and creative domains (Memmert & Bittner, 2022), we leveraged GLM to assist humans in creatively addressing open-ended, ill-structured problems through DSR. The experts' positive feedback suggests that such support could be helpful, particularly because it can provide humans with inspiration and new ideas to extend and sharpen their concepts and with pointers to more effectively gather information, matching DSR's iterative nature. Such CLS could be considered conceptual scaffolding, helping students becoming aware of knowledge gaps through cues (M. C. Kim & Hannafin, 2011). In challenging the students (Wood et al., 1976) to explore new ideas by offering additional perspectives, it might also fulfill the *strategic scaffolding* purpose (M. C. Kim & Hannafin, 2011). Particularly for students who have nobody around to challenge their ideas, this might be helpful, reducing the chance of "marrying the first ideas" (E5). We use scaffolding modes like hints or expert modeling (e.g., rephrasing concept elements) known to be effective (N. J. Kim et al., 2018). However, students must learn treating suggestions as – potentially false (Lin et al., 2022) – hypotheses requiring validation.

Besides the canvas (meta-level support), CLS might further reduce "the number of constituent acts required to reach solution" (Wood et al., 1976, p. 98), by offering conceptual elements as a basis for iteration (E10) as opposed to an empty sheet. As people might lose motivation when overwhelmed (Corbalan et al., 2008), reducing complexity might lead to students retaining an interest, fulfilling the *direction maintenance* scaffolding function (Wood et al., 1976). As motivation increases if students self-select to add scaffolding (N. J. Kim et al., 2018), we have students themselves request suggestions.

Based on the theoretical perspective, we assume the tool to be more helpful the more severe the issues are for a particular student. Even still, multiple experts explained that beyond novices, even experts could benefit from AI-based scaffolding. However, a certain understanding of DSR and the domain is required. Future research should investigate the relative importance of these issues and DSR proficiency.

The canvas reflects the problem and solution space through lists of concept elements, iteratively developed by the learner, supported by AI-generated suggestions for adjusting or extending each of the lists. Though experts expect benefits from CLS for both problem and solution space exploration – even suggesting broadening the scope – the effectiveness requires further investigation.

Having developed our tool deductively from the issues of novices solving problems, it might be adaptable to other creative problem-solving approaches, such as design thinking (proposed by interviewee). Our approach of generating dynamic scaffolding is flexible, requires no additional training data, and is adjustable to other structures (e.g., canvases). We thus believe such a GLM-based approach could be a fruitful avenue for research on AI-generated scaffolding for problem-solving.

The tool might take over the scaffolding function of *demonstrating solutions*, expecting the "learner will then 'imitate' it back in a more appropriate form" (Wood et al., 1976, p. 98). Several interviewees, however, raised concerns about undesired use patterns. Instead of using suggestions as inspiration to challenge their ideas, students might copy them without reflection. This, though, is not a concern specific to *generative AI*. In AI-assisted decision-making, the phenomenon of overreliance is well-known (e.g., Bućinca et al., 2021). Similarly, in the design area, the phenomenon of "design fixation" (Jansson & Smith, 1991) refers to the effect of humans being less creative when shown solution examples. As DSR is an approach for *designing* artifacts, and AI suggestions could constitute (partial) solution examples, design fixation might occur. Therefore, we propose to explore using AI-assisted decision-making and design fixation knowledge.

Our design follows Law et al. (2020) in providing holistic scaffolding combining hard (template) and soft, dynamic scaffolding (CLS). We emphasize CLS, as there already is structural, tool-based facilitation available, but suggest CLS should complement meta-level facilitation. The tool is not intended to replace teaching DSR but being used along a DSR course or during final thesis preparation, allowing the supervisor to focus on higher-level feedback.

17.6.3 Limitations

We did not derive the canvas design but adapted an existing approach, as developing a DSR canvas is beyond the scope of this paper, and, more importantly, the canvas is not our core proposition. On the contrary, our approach, though exemplified along this specific canvas, is flexible and should be tested with other (DSR) canvases. Additionally, for our prototype, we did not derive design features but used an illustrative design as our goal was to get feedback on the general idea of AI-generated, content-level scaffolding. Our tool is not yet implemented technically and has not been tested with end-users. However, our conceptual approach, including generating AI suggestions, can already be executed manually, serving as a foundation for the next iteration. According to scaffolding literature and experts interviewed, positive effects of using the tool can be expected. However, these should be measured in an end-user study.

17.6.4 Contribution and Outlook

We showcase a flexible way of combining static and dynamic scaffolding with novel technology. We deductively developed DPs, instantiated them as illustrative tool, and evaluated them through expert interviews. We thereby contribute design knowledge in the form of a situated implementation of the artifact (level 1) as well as one aspect toward a nascent design theory (level 2), according to Gregor and Hevner (2013). Through this new focus on content, we complement existing DPs for DSR tool support focusing on meta-level support (e.g., vom Brocke et al., 2017).

We show that AI-generated, content-level scaffolding might be a pathway to supporting students. For the next iteration, we suggest a naturalistic evaluation (Venable et al., 2016). We offer ideas for design improvements, particularly for the human-AI interaction. Going forward, integrating AI-generated, content-level scaffolding into DSR teaching should be explored. Regarding the challenge of human-AI collaboration for problem-solving (Dellermann et al., 2019), we propose investigating *generative* AI. We offer a structured approach to generating suggestions along an existing solution framework (i.e., canvas), potentially adaptable to other creative problem-solving approaches.

17.7 Conclusion

Students must experience DSR by applying it to real-world problems to learn DSR effectively. We propose an illustrative tool to support them, as navigating the new paradigm and a new domain can be difficult. Conceptualizing DSR as a problem-solving approach, we use scaffolding as a theoretical and GLM as a technical foundation. We show that complementing (partially existent) meta-level support with content-level support is perceived as helpful. Proposing design knowledge through DPs and an illustrative, situated instantiation, we directly contribute towards an ecosystem of DSR tool support, as suggested by Morana, vom Brocke, et al. (2018) and show that, as suggested by Doo et al. (2020), AI-based scaffolding can be used in (higher) education. More broadly, due to our deductive approach, we contribute to the understanding of how humans can be supported by AI to creatively solve problems, which was described as a challenge for human-AI collaboration (Dellermann et al., 2019). We encourage our community to further explore generative AI like GPT-3 or ChatGPT for this challenge.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

17.8 References

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions. In *48th Hawaii International Conference on System Sciences*, USA.

- Contell, J. P., Díaz, O., & Venable, J. R. (2017). DScaffolding: A Tool to Support Learning and Conducting Design Science Research. In A. Maedche, J. vom Brocke, & A. Hevner (Eds.), *Lecture Notes in Computer Science. Designing the Digital Transformation* (Vol. 10243, pp. 441–446). Springer International Publishing. https://doi.org/10.1007/978-3-319-59144-5_28
- Corbalan, G., Kester, L., & van Merriënboer, J. J. (2008). Selecting Learning Tasks: Effects of Adaptation and Shared Control on Learning Efficiency and Task Involvement. *Contemporary Educational Psychology*, 33(4), 733–756. <https://doi.org/10.1016/j.cedpsych.2008.02.003>
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Doo, M. Y., Bonk, C., & Heo, H. (2020). A Meta-Analysis of Scaffolding Effects in Online Learning in Higher Education. *The International Review of Research in Open and Distributed Learning*, 21(3), 60–80. <https://doi.org/10.19173/irrodl.v21i3.4638>
- Fischer, A., Greiff, S., & Funke, J. (2012). The Process of Solving Complex Problems. *The Journal of Problem Solving*, 4(1). <https://doi.org/10.7771/1932-6246.1118>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Gau, M., Maedche, A., & vom Brocke, J. (2022). DSR Buddy: A Conversational Agent Supporting Design Science Research Activities. *Wirtschaftsinformatik 2022 Proceedings*, 35. https://aisel.aisnet.org/wi2022/prototype_track/prototype_track/35/
- Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *38th International Conference on Information Systems*. <https://aisel.aisnet.org/icis2017/HCI/Presentations/1>
- Goldkuhl, G., Ågerfalk, P., & Sjöström, J. (2017). A Design Science Approach to Information Systems Education. In A. Maedche, J. vom Brocke, & A. Hevner (Eds.), *Lecture Notes in Computer Science. Designing the Digital Transformation* (Vol. 10243, pp. 383–397). Springer International Publishing.
- Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, 30(3), 611–642. <https://doi.org/10.2307/25148742>
- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Grévisse, C., Rothkugel, S., & Reuter, R. A. P. (2019). Scaffolding Support Through Integration of Learning Material. *Smart Learning Environments*, 6(1). <https://doi.org/10.1186/s40561-019-0107-0>
- Hansen, H., & Hebart, M. N. (2022). Semantic features of object concepts generated with GPT-3. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

- Hevner, A. R. (2021). Pedagogy for Doctoral Seminars in Design Science Research. In S. Aier, P. Rohner, & J. Schelp (Eds.), *Engineering the Transformation of the Enterprise* (pp. 185–198). Springer International Publishing. https://doi.org/10.1007/978-3-030-84655-8_12#DOI
- Hevner, A. R., Chatterjee, S., Gray, P., & Baldwin, C. Y. (2010). *Design research in information systems: Theory and practice. Integrated series in information systems: Vol. 22*. Springer.
- Jansson, D. G., & Smith, S. M. (1991). Design Fixation. *Design Studies*, 12(1), 3–11. [https://doi.org/10.1016/0142-694X\(91\)90003-F](https://doi.org/10.1016/0142-694X(91)90003-F)
- Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J. (2022). PromptMaker: Prompt-based Prototyping with Large Language Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems* (pp. 1–8). ACM. <https://doi.org/10.1145/3491101.3503564>
- Karlsson, L., Koivula, L., Ruokonen, I., Kajaani, P., Antikainen, L., & Ruismäki, H. (2012). From Novice to Expert: Information Seeking Processes of University Students and Researchers. *Procedia - Social and Behavioral Sciences*, 45, 577–587. <https://doi.org/10.1016/j.sbspro.2012.06.595>
- Kim, M. C., & Hannafin, M. J. (2011). Scaffolding Problem Solving in Technology-Enhanced Learning Environments (TELEs): Bridging Research and Theory with Practice. *Computers & Education*, 56(2), 403–417. <https://doi.org/10.1016/j.compedu.2010.08.024>
- Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of Computer-Based Scaffolding in the Context of Problem-Based Learning for Stem Education: Bayesian Meta-analysis. *Educational Psychology Review*, 30(2), 397–429. <https://doi.org/10.1007/s10648-017-9419-1>
- Knauss, E. (2021). Constructive Master's Thesis Work in Industry: Guidelines for Applying Design Science Research. In *43rd International Conference on Software Engineering* (pp. 110–121). <https://doi.org/10.1109/ICSE-SEET52601.2021.00021>
- Kuechler, W., & Vaishnavi, V. (2012). A Framework for Theory Development in Design Science Research: Multiple Perspectives. *Journal of the Association for Information Systems*, 13(6), 395–423. <https://doi.org/10.17705/1jais.00300>
- Law, V., Ge, X., & Huang, K. (2020). Understanding Learners' Challenges and Scaffolding their Ill-structured Problem Solving in a Technology-Supported Self-Regulated Learning Environment. In M. J. Bishop, E. Boling, J. Elen, & V. Svihla (Eds.), *Handbook of Research in Educational Communications and Technology* (pp. 321–343). Springer International Publishing. <https://doi.org/10.1007/978-3-030-36119-814>
- Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *CHI'22: Conference on Human Factors in Computing Systems*.

- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Mayring, P. (2014). *Qualitative Content Analysis - Theoretical Foundation, Basic Procedures and Software Solution*.
- Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>
- Memmert, L., Cvetkovic, I., & Bittner, E. (2023). Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/405
- Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. *Journal of Information Systems Education*, 34(3), 279–292. <https://aisel.aisnet.org/jise/vol34/iss3/3>
- Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the Association for Information Systems*, 16(9), 799–837. <https://doi.org/10.17705/1jais.00408>
- Meuser, M., & Nagel, U. (2009). Das Experteninterview — Konzeptionelle Grundlagen und Methodische Anlage. In S. Pickel, G. Pickel, H.-J. Lauth, & D. Jahn (Eds.), *Methoden der vergleichenden Politik- und Sozialwissenschaft* (pp. 465–479). VS Verlag für Sozialwissenschaften.
- Morana, S., Scheid, M., Gau, M., Benke, I., vom Brocke, J., Fettke, P., & Maedche, A. (2018). Research Prototype: The Design Canvas in MyDesignProcess.com. In *13th International Conference on Design Science Research in Information Systems and Technology*.
- Morana, S., vom Brocke, J., Maedche, A., Seidel, S., Adam, M. T. P., Bub, U., Fettke, P., Gau, M., Herwix, A., Mullarkey, M. T., Nguyen, H. D., Sjöström, J., Toreini, P., Wessel, L., & Winter, R. (2018). Tool Support for Design Science Research—Towards a Software Ecosystem: A Report from a DESRIST 2017 Workshop. *Communications of the Association for Information Systems*, 43, 237–256. <https://doi.org/10.17705/1CAIS.04317>
- Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., Kyza, E., Edelson, D., & Soloway, E. (2004). A Scaffolding Design Framework for Software to Support Science Inquiry. *Journal of the Learning Sciences*, 13(3), 337–386. https://doi.org/10.1207/s15327809jls1303_4

- Reiser, B. J. (2004). Scaffolding Complex Learning: The Mechanisms of Structuring and Problematising Student Work. *Journal of the Learning Sciences*, 13(3), 273–304. https://doi.org/10.1207/s15327809jls1303_2
- Saye, J. W., & Brush, T. (2002). Scaffolding Critical Reasoning About History and Social Issues in Multimedia-supported Learning Environments. *Educational Technology Research and Development*, 50(3), 77–96. <https://doi.org/10.1007/BF02505026>
- Sweller, J., & Candler, P. (1992). The Split-attention Effect as a Factor in the Design of Instruction. *British Journal of Educational Psychology*, 62(2), 233–246. <https://doi.org/10.1111/j.2044-8279.1992.tb01017.x>
- Thuan, N. H., & Antunes, P. (2022). Positioning Design Science as an Educational Tool for Innovation and Problem Solving. *Communications of the Association for Information Systems*, 51. <https://aisel.aisnet.org/cais/vol51/iss1/1>
- Thuan, N. H., Drechsler, A., & Antunes, P. (2019). Construction of Design Science Research Questions. *Communications of the Association for Information Systems*, 332–363. <https://doi.org/10.17705/1CAIS.04420>
- van Merriënboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking the Load Off a Learner's Mind: Instructional Design for Complex Learning. *Educational Psychologist*, 38(1), 5–13. https://doi.org/10.1207/S15326985EP3801_2
- Venable, J. (2006). The Role of Theory and Theorising in Design Science Research. In *1st International Conference on Design Science in Information Systems and Technology*.
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, 25(1), 77–89. <https://doi.org/10.1057/ejis.2014.36>
- vom Brocke, J., Fettke, P., Gau, M., Houy, C., & Morana, S. (2017). Tool-Support for Design Science Research: Design Principles and Instantiation. *SSRN Electronic Journal*. Advance online publication. <https://doi.org/10.2139/ssrn.2972803>
- vom Brocke, J., Winter, R., Hevner, A., & Maedche, A. (2020). Special Issue Editorial – Accumulation and Evolution of Design Knowledge in Design Science Research: A Journey Through Time and Space. *Journal of the Association for Information Systems*, 21(3), 520–544. <https://doi.org/10.17705/1jais.00611>
- Winter, R., & vom Brocke, J. Teaching Design Science Research. In *ICIS 2021 Proceedings*. <https://aisel.aisnet.org/icis2021/pdw/pdw/4>
- Wood, D., Bruner, J. S., & Ross, G. (1976). The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>

17.9 Author Biographies

Lucas Memmert is a research associate and PhD candidate at University of Hamburg, Germany. He is part of the research group Information Systems and Socio- Technical Design. In his research he focuses on human-AI collaboration, specifically on humans and generative AI systems collaborating to solve creative and complex problems. He is part of the research group “Hybridizing human and artificial intelligence in knowledge work (HyMeKI)”, a project funded by the German Federal Ministry of Education and Research (BMBF).



Navid Tavanapour is a postdoc and member of the research group Information Systems, Socio-Technical Design at the University of Hamburg, Germany. He is specialized in the research areas of human-AI-collaboration. In this context, he investigates designs and develops solutions for the interplay between intelligent technologies with individuals in various domains. His research also includes collaborative systems, AI-support, conversational AI, and hybrid intelligence and has been published in journals such as *Information Systems Frontiers*, *Informatik Spektrum* and *AIS Transactions on Human-Computer Interaction*. He presented his research at conferences such as the *International Conference on Information Systems*, *European Conference on Information Systems*.



Eva Bittner is a Professor for Information Systems and Socio- Technical Design at University of Hamburg, Germany. She is also head of the junior research group “Hybridizing human and artificial intelligence in knowledge work (HyMeKI)” and a member of Germany’s Platform Learning Systems expert group on the future of work and human-machine interaction. Her research focuses on understanding and designing collaborative work practices and socio technical systems for knowledge intense work. Her research has been published in international journals such as the *Journal of Management Information Systems (JMIS)*, *Business and Information Systems Engineering (BISE)*, *Information and Management (I&M)*, *Information Systems Frontiers (ISF)* and presented at international conferences, e.g. the *International Conference on Information Systems (ICIS)* and the *European Conference on Information Systems (ECIS)*.



17.10 Appendix - Interview Questions

Experience

- How did you come in contact with DSR novices?
 - Do your students use DSR in term papers or final theses?

- Have you taught DS(R) courses?

Overarching structure

- How does the overarching structure support students in developing an understanding for how to develop prototypes using DSR?
- Do you provide students with a structure?

Content-level support

- How can content-level suggestions for concept elements support students?
- Which changes do you expect with regards to the depth and breadth of development?
- How could the speed towards the first concept draft change?
- Do you support your students on content-level?

Suggestion integration

- What is the impact of having the suggestions displayed directly in the tool?
- How does this influence the need to work with other tools?

Timely information provision

- How could the ways of working of students improve by having the suggestions displayed just below concept elements?

Coverage support

- How can the tool support the development of adequate solutions?
- How can it support achieving a good coverage of the solution with respect to the problem?

Relationship suggestions

- How can suggestions for relationships between concept elements support students?

Formulation support

- How can theory-driven formulation support students?
- Is there an additional benefit for having AI suggestions additionally to the general structure of the concept element (design principle)?

18 Publication 9: Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions

Memmert, L., Cvetkovic, I., & Bittner, E. (2023). Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions. In *Proceedings of the 31st European Conference on Information Systems (ECIS)*, Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/405

Abstract

Solving complex problems was named a new challenge for research on human-AI collaboration. In our study we focus on a particular means to solving complex problems: design science research (DSR). We investigate whether AI, more specifically, generative language models (GLM), can support an individual in conceptualizing DSR studies by making helpful suggestions. To do so we use extracts of a published DSR study and have GPT-3, a GLM provide suggestions for aspects of this study. These suggestions are then evaluated in a survey (n=33) regarding their helpfulness. Results show that GLM suggestions are perceived to be helpful, with some variation depending on expertise. Reported interest in using such a tool in the future was high. Describing how GLMs can offer helpful suggestions we contribute toward a DSR tool support ecosystem and, more generally, towards knowledge on how humans and (generative) AI systems can team up to solve complex problems.

Keywords: GPT-3, language model, design science research, human-AI collaboration

18.1 Introduction

With advances in artificial intelligence (AI) there is an increased research interest in teaming up of AI systems and humans (Seeber, Bittner, et al., 2020). A new challenge is solving complex problems through human-AI collaboration (Akata et al., 2020; Dellermann et al., 2019). In our study, we envision human-AI collaboration for a particular approach to solving complex problems: design science research (DSR).

Despite DSR's importance a lack of tool support for conducting DSR studies was highlighted, accompanied by a call for research to establish a DSR tool support ecosystem (Morana, vom Brocke, et al., 2018). Tools have been proposed for facilitating DSR researchers in different aspects such as collaboration within the team, research process documentation or result communication (e.g., vom Brocke et al., 2017). Some tools leverage intelligent features, e.g., adding knowledge elements via a conversational agent (Gau et al., 2022). In making the vast DSR knowledge more accessible they might particularly benefit novices (Gau et al., 2022). However, these tools focus on supporting researchers on a meta-level.

We investigate how this support on meta-level can be complemented through support on a content-level. More specifically, we propose to imagine an AI-based tool which can be used by individual researchers when developing a concept for a DSR study. Throughout the DSR project, with the human incrementally building and adjusting the concept, the tool

would make suggestions on different aspects of the concept to inspire the human to explore the problem and solution space (vom Brocke et al., 2020), and the in-between relationships (Venable, 2006) more comprehensively. Such suggestions would have to be treated with caution and would have to be rigorously grounded by the researcher before inclusion into the concept. Thus, instead of providing general guidelines, e.g., on how to formulate a design principle (i.e., meta-level support), the tool would suggest an actual, study-relevant design principle, based on the information the user has already entered (i.e., content-level support). The user could generate new suggestions to receive further inspiration. Thereby, the user and the AI would iteratively work towards the joined goal (DSR concept) by adding and adjusting design components, considering each other's contributions (i.e., *collaboration*, Bedwell et al., 2012). To aid researchers, the tool should provide suggestions for the specific study, inspiring them to enhance the concept. As DSR studies are concerned with novel problems or solutions to make a contribution (Gregor & Hevner, 2013), creating a pool of potential suggestions beforehand for all the studies different user might want to conduct seems difficult.

Instead of creating a pool of potential suggestions for the AI to select from, we use a *generative language model* (GLM). GLM have been shown to be able to *generate* context-specific text based on a given input (Brown et al., 2020), eliminating the problem of having to create a pool of potential suggestions beforehand. In this study we investigate how suggestions generated by 'GPT-3' (Brown et al., 2020), a state-of-the-art language model, are perceived, seeking to answer the following research question:

RQ 1: To what extent are GLM suggestions perceived as helpful for conceptualizing DSR studies?

Similarly to Gau et al. (2022) we assume that particularly novices might benefit from DSR tool support. From complex problem solving literature (psychology) it is known that novices find it more difficult to explore the problem and solution space appropriately and to build viable internal representations due to their lack of prior knowledge (Fischer et al., 2012). They therefore might find AI suggestions to be more helpful. As perceived helpfulness affects the acceptance of AI and the potential for collaboration success, we investigate the existence of individual differences in the perception of AI suggestions' helpfulness based on level of expertise. We thereby answer the following research question:

RQ2: How does the level of expertise influence the perceived helpfulness?

To answer these research questions we use the GLM 'GPT-3' (Brown et al., 2020) to create AI suggestions. We then survey humans with different levels of expertise on their perception of the helpfulness of those AI suggestions. The results of our analysis show that the overall perception of the AI suggestions is positive. Significant differences in the perceived helpfulness based on expertise are sparse. Opposite from what we expected, higher levels of expertise occur with higher perceived helpfulness ratings. With our paper we contribute to addressing the call for research on a DSR tool ecosystem (Morana, vom Brocke, et al., 2018), demonstrating that the existing facilitative, meta-level support can be complemented

by content-level support. With our innovative approach of generating helpful AI suggestions along a canvas-like structure, we show a potential pathway for humans and AI systems, more specifically GLMs, to collaboratively design solutions for complex problems, an important research challenge (Dellermann et al., 2019).

18.2 Background

18.2.1 Design science research tool support

Solving complex problems was named a new challenge for human-AI collaboration (Dellermann et al., 2019). An approach for solving complex problems is DSR. DSR refers to “a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artifacts, thereby contributing new knowledge to the body of scientific evidence” (A. R. Hevner et al., 2010). Such innovative artifacts might include, e.g., software prototypes. A lot of literature is available including overarching DSR approaches, types of theoretical contributions through DSR or guidelines on formulating meta requirements (e.g. Chandra et al., 2015; Gregor & Hevner, 2013; Hevner et al., 2004). To aid researchers in rigorously applying DSR, tool support was suggested (Morana, vom Brocke, et al., 2018). Proposed solutions focus on facilitating researchers in conducting DSR studies including aspects like team member collaboration, improvement of documentation, or context-independent literature suggestions (e.g., Morana, Scheid, et al., 2018) and even included conversational agents to support researchers in documenting studies (e.g., Gau et al., 2022). In summary, the tools proposed facilitate researchers on a meta-level, e.g., by providing a template for a “good” way of formulating meta requirements, or suggesting established approaches for conducting DSR studies, sometimes even by leveraging AI.

Here, we propose a complementary approach to support researchers on a second layer: on content-level (see table 1). Instead of offering information of how to do something in general, e.g., on how to formulate a meta requirement, a tool should suggest actual contents like issues, design requirement, design principles, or kernel theories for the study at hand, enhancing the concept by inspiring the researcher. Our approach uses a GLM to offer tailored suggestions based on the specific study’s content.

Table 1. DSR tool support overview: meta-level vs. content-level support

Level	Explanation	Illustrative examples of support for the exemplary study: “Develop a design for a conversational agent to facilitate contributors in generating elaborate ideas on idea platforms”	Tool examples
Meta-level	Support independent of the specific study	- List DSR approaches, e.g., three cycle approach by (A. Hevner, 2007), DSR process model by (Peffer et al., 2007) - List reference literature, e.g., for phrasing for design requirements as proposed by (Chandra et al., 2015)	- MyDesignProcess.com (vom Brocke et al., 2017) - DScaffolding (Contell et al., 2017)
Content-level	Support for the specific study at hand	Inspirational suggestions for design requirements, e.g., - Conversational agents should elicit ideas by asking open-ended questions. - Conversational agents should encourage elaboration of ideas.	<Our tool>

18.2.2 Generative language models

GLM are large language models trained on a large corpus of text to generate free form text. They are trained with the goal to predict the next word given a certain input, i.e., given a certain context. Though they were trained task-agnostic only with this general goal, GLM have been shown to perform well on typical natural language processing tasks like translation or question answering and even on generating news articles based on a headline (Brown et al., 2020). They do so without task-specific training, i.e., without a task specific training data set or task-specific updates of the model (Brown et al., 2020). The interaction with the model (Brown et al., 2020) is performed via free-form *text-in*, *text-out*. The input might, e.g., consist of an instruction and/or exemplary pairs of inputs and expected outputs (e.g., ‘Italy -> Rome’), followed by another input (‘France ->’), for which the model then generates an output (Paris) accordingly, potentially “recognizing” the presence of pairs of countries and their capitals from the corpus used for training. The exemplary input-output-pairs are referred to as *demonstrations*. They can be distinguished from *training examples as demonstrations* do not lead to model weight updates, i.e., are used as additional input during inference time, whereas training examples typically are used to fine-tune models for a specific task including model (weight) updates (Brown et al., 2020). Depending on the number of demonstrations provided during inference time one can distinguish zero-, one- or few-shot usage (0, 1 or multiple demonstrations respectively) (Brown et al., 2020). Performance differences between the approaches vary by task (Brown et al., 2020), but even without demonstrations good results can be achieved. As the AI system’s output is conditioned on the input provided during inference time (and inference parameters, e.g., randomness), the input needs to be crafted carefully. Techniques to shape the input have been developed as part of *prompt engineering research* (e.g., Mishra et al., 2022).

The utility of humans collaborating with generative AI systems was demonstrated in a variety of creative contexts (Memmert & Bittner, 2022) including joint music composition (Suh et al., 2021) or drawing (Zhang et al., 2021). Several GLM-based tools have been

proposed or used to provide inspiration to humans when solving problems, e.g., an “Idea Machine”, supporting humans in idea generation through theory-guided suggestions by an GLM (Di Fede et al., 2022). Gero et al. (2022) show that GLMs can provide “sparks”, i.e., short sentences, described by users to offer “inspiration” and “external perspectives”. Even more so, Zhu and Luo (2022) report that “GPT can perform conceptual design tasks with a reasonable level of competence” when using it develop design suggestions for actual design challenges. Given that previous literature suggests that generative AI/GLMs can support creative activities through suggestions, we investigate whether such GLM suggestion can be helpful to solve problems tackled via DSR projects which are novel, complex, open-ended and knowledge rich (Hevner et al., 2004; Morana, vom Brocke, et al., 2018) requiring creativity for designing solutions (A. R. Hevner et al., 2010).

Content-level support might have been difficult to achieve with more traditional approaches such as recommender systems, as creating a pool of all potential suggestions for all the possible studies the different users might want to conduct (from which an AI system could then choose and recommend suggestions to the human) does not seem feasible. Thus, instead of trying to create a repository of potential AI suggestions and training a recommender system, we instead suggest to use GPT-3 (Brown et al., 2020), a state-of-the-art GLM. The GLM is trained on a large, general corpus and is therefore flexible to *generate* suggestions regarding a variety of topics based on the human input. As a result, neither for training nor for inference a predefined set of potential AI suggestions is required. Given the potential of GLM for offering helpful suggestions for creative problem solving combined with the flexibility due to the on-demand *generation* of suggestions (instead of prior preparation), we thus believe GLMs might be suited for our case of providing suggestions and thereby content-level support for DSR.

18.3 Methodology

Our underlying idea was to support researchers in developing new DSR concepts with the help of AI suggestions. The human would fill in some general context information and the general problem, as well as first ideas. The system would make suggestions based on those ideas, which the user would review and might or might not find helpful to further develop the concept. As our goal was to understand the helpfulness of the AI suggestions and not to evaluate an experimental tool design, we choose to evaluate these suggestions independently to remove the potential influence of the experimental design on the results. Thus, for our study, we extracted contents from an existing, published DSR paper. We then used these contents from the study as input for the GLM in order to retrieve completions, hereafter referred to as ‘suggestions’. In a survey we asked participants to rate the helpfulness of these suggestions according to a measurement instrument we reused from Rhys Cox et al. (2021).

18.3.1 Study approach

In our approach (figure 2) we used a foundational DSR structure (step 1) to develop prompt template strings (step 2), which we then populated with the contents of an existing study (steps 3 & 4) to generate a pool of AI suggestions (step 5) from which we select suggestions (step 6) for our survey (step 7).

Step 1. As a foundation, we used an (exemplary) canvas-like DSR structure (figure 2, #1). We derived this structure from the approach of using components like issues, theories, design requirements, design principles to describe the concept's cornerstones (in different variations e.g., available in Meth et al., 2015 and Gnewuch et al., 2017). We have enriched this canvas by adding situation, general problem, and artifact class, information typically presented in the studies' manuscript. Besides providing a more comprehensive view of the study to humans, this also allowed us to capture the studies broader context, enriching the information, which can be provided to the AI system.

Step 2. To receive context-dependent suggestions by the AI systems it needs to be provided with appropriate (context-dependent) input. As we sought to develop a re-producible and re-usable approach that would work across study contexts, we developed a fixed structure, i.e., prompt templates, which could then be populated with study-specific contents. The template strings contained structuring elements and placeholders for the specific contents of the study as well as a task description. As concepts like "design requirements" or "design principles" carry different meanings depending on context, we included DSR specific explanations according to (Meth et al., 2015).

For designing the template strings, we used prompt design techniques to improve suggestion quality. In particular we used "decomposition reframing" and "itemization reframing" (Mishra et al., 2022). "Decomposition reframing" follows the idea of splitting a larger, more complex tasks into multiple sub-tasks. We applied this technique by identifying the individual design components for which we sought suggestions. Thus, we had suggestions for four smaller prompts (one for each design component) instead of having one large prompt to receive suggestions for the entire study concept. Additionally, we used "itemization reframing", e.g., instead of having a longer paragraph with the different issues, we created an enumerated list in order to "signal" to the GLM that we expect additional issues as a completion. We included placeholders for design components that might have already been created by the user to act as *demonstration examples* (Brown et al., 2020). However, besides information for the situation and general problem such initial user contributions are optional. In total, we developed four prompt templates, one for each design component. As in DSR, the design components are related or derived from each other (Gnewuch et al., 2017), we included design components of previous stages into the prompt for the next type of design components to increase internal consistency, e.g., the prompt template for design

requirements included placeholders for issues and kernel theories and the template for design principles included placeholders for design requirements and kernel theories (see figure 1#2).

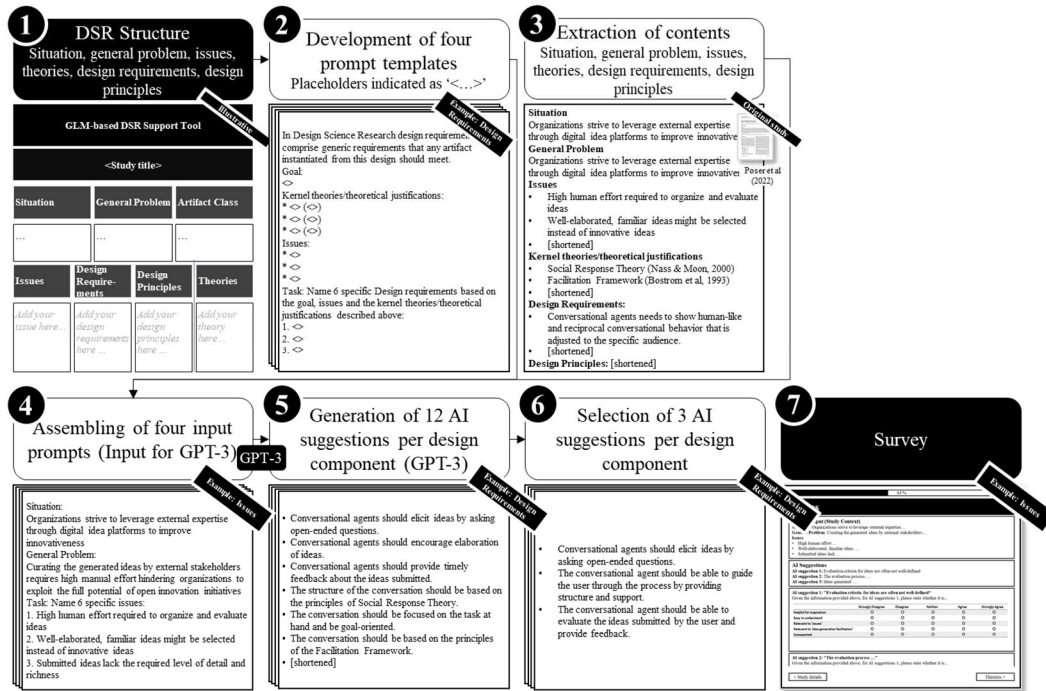


Figure 1. Study approach: from generic DSR “canvas” to context-specific survey contents with GPT-3 suggestions (steps 2-7 contain the actual data)

Step 3. We then populated our template strings with actual contents of a published DSR study. We therefore selected a *published* DSR article (“Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes”, Poser et al., 2022) in order to ensure the foundational study idea and the resulting AI suggestions were based on a paper idea that was both interesting and relevant. Our approach, however, is general, thus any other DSR study could have been selected (or can be selected by other authors). We selected a paper published after the training data collection for the corpus for the GLM we use was completed to ensure no training data leakage occurred, i.e., to avoid the possibility of the study being part of the training data corpus. This was necessary as Carlini et al. (2023) showed that GLM might reproduce text seen during training in verbatim.

From this study, we extracted the following key conceptual elements: situation, general problem, issues, kernel theories/theoretical justifications, design requirements, and design principles. Some elements like design requirements or design principles were clearly described as such in the original study and were thus quoted directly whereas other elements were paraphrased by the authors of this paper.

Step 4. In order to receive the study-specific suggestions we filled in the placeholders with the study contents that we had extracted from the published study.

Step 5. We then fed the resulting populated template strings as prompts into our GLM, GPT-3, to generate suggestions (see figure 1#5 for actual examples). We used the out-of-the-box GPT-3 model (i.e., without fine-tuning) at pre-defined settings (engine: text-davinci-002; temperature: 0.7).

We imagined a usage scenario as follows: while adding their ideas into the tool the user might invoke the generation of AI suggestions. The user then might check these suggestions and might update their concept, before proceeding to generate new suggestions, iteratively building their concept. For each re-generation the number of new suggestions was limited to three in order not to overwhelm the user. We assumed during typical usage a user might refresh (i.e., re-generate) the suggestions up to three times. Simulating this procedure, we generate 12 suggestions for each of the four design components (i.e., 12 suggestions per design component x 4 design component types = 48 suggestions).

Step 6. We were intrigued by the quality of some of the AI suggestions, while others seemed generic, and some occurred multiple times. Our goal was to understand if the AI system could provide helpful suggestions, not if all suggestions generated by the AI were helpful. Given our goal and the described usage scenario, we decided to select three of the twelve suggestions for each of the four design components, as even this reduced number of suggestions would allow us to understand, if suggestions were perceived to be helpful (and in which ways), but would significantly reduce the burden on participants, potentially increasing the response quality and completion rate.

Step 7. The methodology regarding the survey is described in the next subsection.

18.3.2 Survey

To answer our research questions regarding the helpfulness of the AI suggestion and the influence of expertise we conducted a survey gathering the relevant data and performed the corresponding analyses. Besides demographics, we asked participants to reflect their level of expertise with regard to the DSR paradigm and the topic of the study (independent variables). To determine the expertise with regard to the DSR paradigm we asked participants for how many projects or studies they had used DSR along five categories (never, 1 or 2 times, 3 to 5 times, 5 to 10 times, more than 10 times). The familiarity with regards to the topic of the foundational study (i.e., idea generation facilitation) was rated along a 5-point likert scale ('Not at all familiar' to 'Extremely familiar'). Besides these more direct measures of expertise, we also measured the "general expertise" via experience in academic research, i.e., role in academia, on a 5-category scale (Bachelor student to Professor).

After introducing the foundational study (we provided the abstract of the study and asked participants to read it carefully) we asked participants to score each AI suggestions (dependent variables). To do so we had four pages (one page for every design component: issues, theories, design requirements, and design principles) with a short overview of relevant information we had extracted from the study (e.g., for the issues page we presented a

short description of the situation, the general problem, and 3 issues from the study for comparison) as well as the three AI suggestions (e.g., for the issues page we presented three issues). Participants were then asked to score each AI suggestion individually (see figure 1, #7).

For measuring the helpfulness of the AI suggestions according to RQ 1 we used the instrument (both measurement dimensions and scale) from Rhys Cox et al. (2021). Similar to our study, Rhys Cox et al. (2021) used this instrument to have participants rate cues/prompts offered during ideation. According to this instrument, ratings are gathered along five dimensions: ‘helpful for inspiration’, ‘easy to understand’, ‘relevant to <task>’, ‘relevant to <domain>’, and ‘unexpected’. For simplicity, we refer to those five dimensions collectively as ‘helpfulness’ in this study. In our case, the <task> was to find an appropriate *design component*. The <domain> was the topic of the study, i.e., *idea generation facilitation*. For each dimension participants provide a rating on a 5-point likert scale (strongly disagree to strongly agree, coded as 1 to 5). In total participants thus provided 60 ratings with regard to the AI suggestions: 12 suggestions (4 design components with 3 suggestions each) x 5 evaluation dimensions. At the end of the survey, we asked participants if they were interested in using or testing such an AI-based DSR support tool and if they had suggestions for further design components or features the tool should support. In total, the survey contained 23 questions and required about 15 minutes to complete.

To recruit participants with different levels of experience we reached out to students and participants of DSR university courses as well as authors who published in DSR-related tracks at information systems conferences. As we also asked addressees to forward our survey to students and colleagues the total number of recipients and consequently the response rate is unknown.

In total, we had 33 complete submissions. For the analysis we excluded partial submissions (37.1% completion rate) as most participants (75.1%) who did not complete the survey stopped before making any ratings and average ratings for participants who completed the survey were not significantly different from average ratings of participants who only partially completed it (Kruskal-Wallis: Statistic=2.292, $p=0.130$). Among the 33 complete submissions participants’ (female=6, male=26, other=1; age: min=18, max=51, mean=30.5) levels of expertise were diverse (figure 2).



Figure 2. Expertise: a) role or general expertise, b) DSR expertise, c) topic expertise

In order to investigate the general perceived helpfulness of the AI suggestions (RQ1) we calculated the descriptive statistics of the ratings. We tested if they were significantly greater than the neutral rating and if there were any significant differences in ratings among

design components, among dimensions, and among individual AI suggestions within their groups. To understand the impact of expertise on the ratings (RQ2) we performed a correlation analysis between the types of expertise and the ratings.

18.4 Results

We first ask whether the AI suggestions are perceived as overall helpful (RQ1). As table 2 shows, the answer is yes. On average, AI suggestions are scored at 3.5, meaning between ‘neither’ (3.0, i.e., neutral) and ‘agree’ (4.0). A Wilcoxon signed-rank test (an alternative to one sample t test for data deviating from normality, Shapiro-Wilk $p < .001$) for the alternative hypothesis ‘greater than 3’ revealed that ratings across all design components and dimensions except from the dimension ‘unexpected’ were rated significantly greater than neutral (i.e., 3.0, table 3).

In order to understand differences among design components, dimensions, and individual suggestions, we performed the Kruskal-Wallis test (non-parametric alternative to ANOVA, for data that does not fulfill the assumption of homogeneity of variances, Levene’s test: $p < .001$) and corresponding Dunn’s post hoc comparisons. Regarding design components (rows in table 2), pairwise post-hoc Dunn test with Bonferroni adjustments showed that AI suggestions for issues, design requirements, and design principles were rated significantly better than for theories (0.1, 0.2, and 0.3 units respectively, $p < .005$).

With regard to the dimensions (columns in table 2), AI suggestions were rated significantly worse with regard to ‘unexpectedness’ as compared to the other dimensions (between 0.8 and 1.1 units; $p < .001$). With regard to the individual AI suggestions there were no significant differences in rating between AI suggestions within the same design component group ($p > .05$). On a higher level, Theories show to be the least easy to understand. DRs and DPs show higher ratings on the ‘Helpful for inspiration’ dimension than Issues and Theories, respectively. Another indication for the AI suggestions having been perceived to be helpful was reflected in the answers to a question, which we asked after participants had seen the kinds of AI suggestions the GLM generated for our exemplary study: “Would you be interested to use a such a Design Science Research support tool that provides these kinds of AI suggestions in the future?”. 81.8 % of participants responded with ‘Yes’, only 15.2 % answered ‘No’ (3.0% N/A). Looking at single design components, Issues 1 and 2 (“Evaluation criteria for ideas are often not well-defined”, “The evaluation process is often biased and subjective”) display the lowest ‘Unexpected’ rating. The ratings for Theory suggestion 3 (“Task-technology fit”) were similar across dimensions. DR2 “The conversational agent should be able to guide the user through the process by providing structure and support” displays highest rating for ‘Relevant to DRs’, but lowest on the dimension ‘Unexpected’. DP2 “Make the conversational agent's behavior adjustable to different types of users by taking into account the user's role, knowledge and expertise, to provide an adequate level

of support, and by offering different types of interactions (e.g., more directive or less directive) to users with different levels of experience.” is both the most helpful one as well as the most relevant to DPs. In addition, DP2 is the only design component that positively correlates with all three expertise types on several dimensions.

Table 2. Mean values and standard deviations for participant ratings for AI suggestions per design component (n=33)

Helpfulness dimensions\ AI suggestions for design components	Helpful for design inspiration	Easy to understand	Relevant to <design component>	Relevant to 'idea generation facilitation'	Unexpected	Overall
Issues	3.3 (1.0)	4.0 (0.9)	3.8 (0.8)	3.5 (0.8)	2.5 (1.0)	3.4 (1.1)
1	3.4 (0.8)	4.0 (1.0)	3.8 (0.8)	3.5 (0.7)	2.3 (0.9)	3.4 (1.0)
2	3.3 (1.2)	3.9 (1.1)	3.7 (1.1)	3.4 (1.0)	2.3 (1.0)	3.3 (1.2)
3	3.2 (1.2)	4.2 (0.8)	3.9 (0.7)	3.5 (0.8)	2.8 (1.1)	3.5 (1.0)
Theories	3.4 (0.9)	3.4 (0.8)	3.5 (0.9)	3.2 (0.8)	3.1 (1.0)	3.3 (0.9)
1	3.3 (0.7)	3.1 (0.9)	3.4 (0.8)	3.1 (0.8)	3.3 (1.0)	3.3 (0.9)
2	3.5 (0.9)	3.6 (0.7)	3.5 (1.0)	3.4 (1.0)	2.9 (0.9)	3.4 (0.9)
3	3.3 (1.1)	3.6 (1.0)	3.5 (0.9)	3.2 (0.7)	3.0 (1.0)	3.3 (1.0)
Design requirements (DRs)	3.8 (0.8)	4.1 (0.9)	3.7 (1.0)	3.7 (0.9)	2.4 (0.9)	3.5 (1.1)
1	3.8 (0.8)	4.2 (0.9)	3.6 (1.1)	3.8 (0.8)	2.5 (1.0)	3.6 (1.1)
2	3.6 (0.9)	4.1 (0.9)	3.8 (1.0)	3.6 (0.9)	2.2 (0.9)	3.5 (1.1)
3	3.8 (0.8)	4.1 (0.9)	3.6 (1.0)	3.7 (0.9)	2.6 (0.8)	3.6 (1.0)
Design principles (DPs)	3.9 (0.8)	3.7 (0.9)	3.9 (0.8)	3.7 (0.7)	2.7 (1.0)	3.6 (0.9)
1	3.7 (0.9)	3.5 (1.1)	3.8 (0.9)	3.8 (0.8)	2.8 (1.0)	3.5 (1.0)
2	4.1 (0.7)	3.8 (0.8)	4.2 (0.6)	3.7 (0.6)	2.5 (0.9)	3.7 (0.9)
3	3.9 (0.7)	3.8 (0.7)	3.7 (0.8)	3.6 (0.7)	2.7 (1.0)	3.5 (0.9)
Overall	3.6 (0.9)	3.8 (0.9)	3.7 (0.9)	3.5 (0.8)	2.7 (1.0)	3.5 (0.9)

Table 3. Results of Wilcoxon signed-rank test (all ratings, design components, dimensions)

	W	df	P
All individual ratings	842636.000	1979	< .001
Design components			
Issues	56910.000	494	< .001
Theories	41846.500	494	< .001
Design requirements	59506.500	494	< .001
Design principles	53219.000	494	< .001
Dimensions			
Helpful for inspiration	41597.000	395	< .001
Easy to understand	47366.000	395	< .001
Relevant to <design component>	43339.000	395	< .001
Relevant to 'idea generation facilitation'	31046.000	395	< .001
Unexpected	10457.000	395	1.000

Note. For the Wilcoxon sign-ranked test, the alternative hypothesis specifies that the median is greater than 3.

Influence of expertise. Regarding the influence of expertise on perceived helpfulness (RQ2) the results are heterogeneous. We analyzed correlations between the role (general

expertise), topic expertise, and DSR expertise and different ratings for concepts elements averaged as well as separately for each design component. Assumption checks were performed before correlation analysis to determine the normality of the data and choose the according correlation coefficient. According to Shapiro's pairwise normality test, most of the variables did not have normal distribution. Thus, for a more comprehensive comparison, we chose Spearman's correlation coefficient which is more robust. While most of the correlations we calculated between expertise and design components did not yield statistically significant results, some of the aspects did. In total we found 19 significant ($p < 0.05$) correlations between expertise (three types) and design components groups/design components (7/12, respectively) across all five dimensions. Three design component groups correlate with role (Theories: $\rho = .367$ (Relevant to 'idea generation facilitation'); DRs: $\rho = .393$ (Helpfulness for inspiration), $\rho = .413$ (Relevant to 'idea generation facilitation'); DPs: $\rho = .404$ (Helpfulness for inspiration)). Two groups correlate with Topic Expertise (DRs: $\rho = .470$ (Relevant to 'idea generation facilitation'); DPs: $\rho = .500$ (Helpfulness for inspiration), $\rho = .387$ (Relevant to design principles)). All correlations were positive, i.e., higher ratings occurred with higher expertise. Most significant correlations concerned the dimensions 'Helpfulness for inspiration' (8) and 'Relevant to idea generation facilitation' (7). We therefore performed a correlation analysis on these two dimensions and found the correlation to be positive, with strongest correlation across DRs ($\rho = 0.7$, $p < .001$).

Additional design components. In our survey participants rated AI suggestions on four design components (issues, theories, design requirements, design principles) and gained an impression of the capabilities of state-of-the-art GLM suggestions. In an open question we asked participants, which other design components they thought should be supported by such an AI tool. While one participant believed that the "core elements" were covered, other suggested additional design components like design features (as in concrete instantiations of the design principles), semantic connections or mappings between the design components to achieve a "mapping diagram", artifact type (e.g., instantiation, method, framework), types of data collection and validation, highlighting of context dependencies, and relevant literature. Most often, participants mentioned the need for evaluation-related aspects.

Additional potential features. We also asked participants which aspects in the design of a tool providing such AI suggestions they thought to be important. Participants suggested that the tool could support the formulation of design components, e.g., a transfer of 'raw' user-written requirements into design requirements of a 'correct' format, e.g., according to Gregor et al. (2020). They also suggested clustering of design knowledge to increase structure, support in the evaluation of design knowledge as well as support in selecting an appropriate prototyping framework.

Usage process. With regard to the usage process, participants had different ideas like integration into existing tools (e.g., citation manager, word processor). For integration into

word processors context-dependent inline suggestions (“like ‘grammaly’”) were mentioned. To improve the usefulness, more detailed explanation of AI suggestions as well as a summary with all suggestions were proposed. Lastly, a participant stressed the need to carefully design the usage process to mitigate potential negative effects.

18.5 Discussion

Overall, suggestions generated with our approach were rated positively regarding helpfulness both by less and more experienced participants, encouraging us to further explore this path of human-AI collaboration.

18.6 Perception of AI suggestions

Ratings for all design components were higher than neutral for all design components. However, ratings for theories were significantly lower than for the other design components. One interpretation is, that GLMs might be less suitable to propose such abstract ideas like theories. However, such differences need to be interpreted with caution as another explanation could be that theories are in general more difficult to interpret than, e.g., issues, and thus would receive lower ratings, independently of originating from human or AI. Future research should therefore investigate the differences also with regard to the suggestion’s origin. While helpfulness ratings were positive there is still room for improvement. In our study, we used different prompt engineering techniques to increase the AI suggestion quality. However, we did not provide additional demonstration examples from other studies as part of the input prompt during inference time, neither did we fine-tune the model. Brown et al. (2020) have shown that providing additional examples can improve the output quality significantly.

The lack of significant differences in ratings between AI suggestions within their groups might be caused by the preselection step we performed. If we had included all AI suggestions into the survey (or only generated 3 suggestions) this might have been different. However, as discussed in the methodology section, a selection of (3) relevant suggestions from the larger pool of suggestions (12) seemed to be more representative for the usage scenario, allowing us to answer the research questions while lowering the burden on survey participants by reducing the number of required ratings.

Expertise. Results regarding correlation between level of expertise and perceived helpfulness were mixed. Other than we expected novices did not perceive AI suggestions to be more helpful. On the contrary, in cases with significant correlation (particularly ‘relevance to <topic>’ and ‘helpfulness for inspiration’) these correlations indicated higher ratings with higher levels of expertise. It is likely that senior researchers or those who are more familiar with the topic might have a greater appreciation for the AI suggestions. Potentially they find the suggestions easier to mentally integrate and relate them with their existing knowledge. This might offset other effects known from CPS literature (Fischer et al., 2012).

However, these assumptions need to be verified in a future study, potentially with a more fine-grained and objective assessment of the participants' expertise and a more in-depth investigation of the reasoning for the ratings.

Unexpectedness. The 'unexpected'-dimension received significantly lower ratings as compared to the remaining dimensions and was the only dimension with an average score (2.7) below neutrality (3.0). One participant reflected this aspect stating: "at no time was I really surprised by the suggestions and thought they are all things that would probably have come up in any discussions we would have as researchers/ practitioners". While collaboration among multiple humans is common in DSR projects (Sein et al., 2011), other humans might not always be cost-effectively available to have discussions. Thus, an "AI teammate" turning individual work of one human into collaborative work of a hybrid, sociotechnical ensemble, e.g., as described by Dellermaun et al. (2019) might add value. Comparatively low user ratings with regards to unexpectedness are not surprising as "generative language models are trained to match the distribution of content generated by humans" (Brown et al., 2020). However, as we do not seek to replace the human by an AI, but have the AI collaborate with the human, e.g., inspiring a more comprehensive problem and solution space exploration, it might be desirable to increase unexpectedness of suggestions (see design fixation literature, e.g., Sio et al., 2015), while retaining high relevance to task and domain, i.e., design component and topic, respectively. Approaches to be explored might include providing additional examples (sourced from high-quality DSR publications) with a more appropriate level of unexpectedness (Zhu & Luo, 2022) or adjusting model parameters like temperature (i.e., level of randomness). Additionally, while with GPT-3 we used a high-performing state-of-the-art model (Brown et al., 2020), with rapid GLM advancements, our results need to be interpreted carefully and might be considered to show a lower boundary of what will be possible.

18.6.1 Usage of AI suggestion in a DSR support tool

In line with the overall positive ratings for the AI suggestions was the large percentage (81.8%) of participants being interested in using a tool that offers such GLM-based suggestions, with more than half of them (61.1%) being interested in testing such a prototype.

Caution when using suggestions. Though AI suggestions were rated positively they should be used with caution as part of a DSR support tool, in particular given the characteristics of GLM and the required rigor and suggested ethical behavior in (design science) research (Benke et al., 2020; Durani et al., 2021; Myers & Venable, 2014). Potential issues include lack of truthfulness of AI suggestions, lack of evidence for AI suggestions, a danger of plagiarism, a lack of understanding for the meaning, and a lack of novelty. As Lin et al. (2022) have demonstrated, GLM completions are not necessarily truthful, but might reproduce falsehoods and biases. Before including suggestion into the study concept it therefore needs to be carefully assessed to "ensure the quality of the artefact" (Myers & Venable, 2014, p. 806), which should include a reflection on ethical principles (Benke et al., 2020).

Another issue is the lack of evidence or support. DSR design components need to be thoroughly grounded in theory or practice, e.g., based on empirical findings (Goldkuhl, 2004). The AI suggestions provided by the tool, however, are not backed by empirical data or literature and therefore cannot be used directly. They might rather be treated as means for building “hypotheses” that need to be verified. A third issue is the potential danger of plagiarisms. It was shown that GLM might replicate entire passages of text from the underlying training data in verbatim based on certain prompts (Carlini et al., 2023). This, however, is usually not transparent from the output. Using the output as part of a study without proper attribution might constitute plagiarism. A fourth potential issue might be the lack of understanding for the meaning of the suggestions, i.e., the GLM cannot reason about developing the concept like a human but produces texts according to a probability distribution. It cannot make moral judgements (Floridi & Chiriatti, 2020), but these are required to be performed by the human (Durani et al., 2021). Lastly, the lack of unexpectedness we found in our study might appear to be contrary to the goal of DSR of developing novel artefacts. However, as Gregor and Hevner (2013, p. 344) point out, “nothing is really ‘new’”, everything “builds on previous ideas”. Together with the previous issue (lack of understanding), this opens up an interesting research challenge: given the types of theorizing and reasoning in DSR (e.g., Kuechler & Vaishnavi, 2012; Venable, 2006) and the ability of GLMs to mimic “common sense reasoning” (Brown et al., 2020), what types of contributions (e.g., Gregor & Hevner, 2013) can we expect from GLMs?

Sociotechnical design. Given these challenges, a careful, sociotechnical design will be required, to reduce the risk of users only insufficiently investigating a problem and the solution during a DSR study, accepting even unfit AI suggestion. One participant stressed this aspect as follows: “I think you need to carefully design the whole socio-technical usage process to not design a tool that makes inexperienced DSR users to accept useless suggestions.” The lack of critical engagement with AI suggestions resulting in an *overreliance* of humans on AI systems is a known problem from the related field of AI-assisted decision making (e.g., Bućinca et al., 2021). Narrow solution space exploration is also a known problem from the area of design studies referred to as *design fixation*, the phenomenon that humans can be influenced by design (aspects) they have encountered before and might as a result be less innovative (Jansson & Smith, 1991; Sio et al., 2015). In developing socio-technical solutions it might therefore be worthwhile to leverage approaches known from AI-assisted decision making and design studies literature to increase user’s engagement with the AI suggestions and prevent narrow problem and solution space exploration, e.g., by having users actively request information (Bućinca et al., 2021) or by providing less common examples (Sio et al., 2015). Relevance to real-world practice might be considered important for DSR (Gregor & Hevner, 2013). The sociotechnical design should include the tool’s usage processes as part of a DSR project. Future research should explore how such a tool can support establishing relevance, e.g., by supporting empirical data gathering (preparation) or by summarizing already available observations from practice.

Beyond a sociotechnical design, however, it will be important to educate humans to critically reflect on AI suggestions. For our case, users would need to be educated about the limitations of such tools and that all suggestions need to be treated as unsubstantiated hypotheses, not as facts, as well as on ethical approaches towards DSR (e.g., Durani et al., 2021). More broadly, question around ownership, authorship, and attribution of AI outputs will need to be clarified (Stokel-Walker, 2023).

18.6.2 Towards a DSR support tool with embedded, generative AI capabilities

Generative AI has been shown to support humans in creative, open tasks (Jiang et al., 2022; Suh et al., 2021; Zhang et al., 2021). In this study, we show how GLMs can be used to support a human in solving an open, complex problem, exemplified along concept development for designing an artifact according to the DSR paradigm. Our study approach allowed us to evaluate the helpfulness of such GLMs independently from a specific tool implementation. However, future research should explore implementing a prototype (figure 3), potentially using our approach as a foundation. Instead of extracting contents from a published study, the system would use the inputs of the human from the user interface and plug these into the template strings we have developed (step 2) to generate AI suggestions, which are then displayed to the human. With such a prototype, future studies should explore the validity across topics (select a different foundational study at step 3) or the influence of different prompt engineering techniques (adjust templates in step 2) to better understand the potentials and limitations of such an approach.

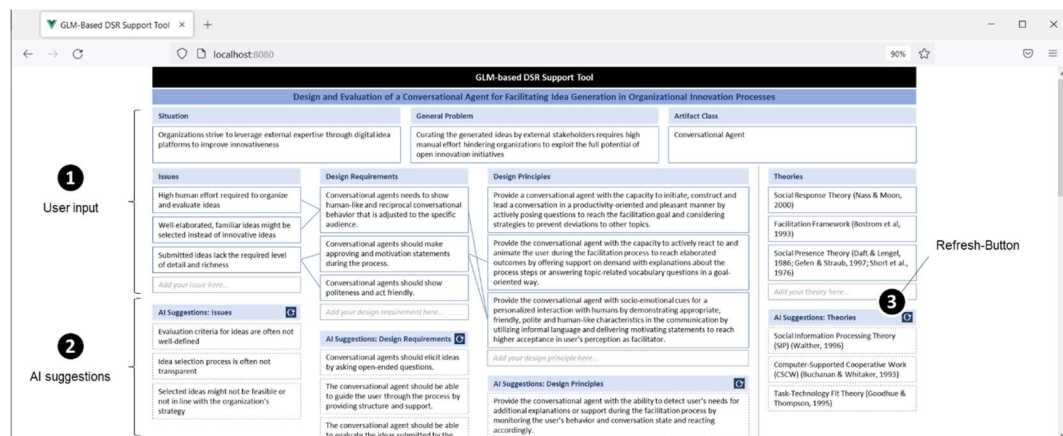


Figure 3. Prototype idea for a GLM-based DSR support tool based on our study approach: 1) user input area, 2) AI suggestion area, 3) refresh-buttons to generate new suggestions for each design component (illustrative contents for demonstration cited or paraphrased from Poser et al., 2022)

With this template-based approach, we simplify the interaction of the human with the GLM, as the inputs for the GLM (i.e., prompts) are automatically assembled in the background and the users can focus on concept development. Through this approach, generative AI is embedded into the tool and users do not have to learn to directly interact with this new technology (i.e., GLM) before working with the tool. This approach is in alignment

with the challenges highlighted by Jiang et al. (2022). In their study, participants interacting with a GLM “quickly noticed that different paraphrases of the same prompt could have vastly different levels of effectiveness towards producing the desired effect”. Jiang et al. (2022) state that “[...] beginners were particularly surprised when their initial zero-shot prompts did not behave as intended” and often relied on “acquiring tips and tricks”. With our approach we abstract away such technology-specific aspects, allowing even non-tech-savvy users to leverage GLM capabilities while focusing on designing their artifact. Future research should explore how such human-AI collaboration affects work performance (Rafner et al., 2021), i.e., DSR concept (component) quality. More broadly, the question arises how such generative AI embedded into a tool compares to (conversational) free-text interaction with GLMs (e.g., ChatGPT). We argue that for our case, embedding generative AI could allow users to focus on concept development, without having to learn effective prompting strategies, e.g., as part of the pre-defined prompts in the tool established practices for phrasing design requirements or design principles (e.g., Chandra et al., 2015) that might be unknown to (novice) users can be included.

In this study, we focus on designing DSR solutions along a DSR canvas. However, we suggest future research to explore if other open, complex problems can be supported in a similar fashion. In alignment with our approach an established structure (figure 2, #1) like a canvas (e.g., business model canvas) could be a potential starting point for prompt template development which includes context-specific definitions of terms (#2). Such research could explore the transferability of the approach onto other complex problem solving mechanisms.

In summary, while initial AI suggestions quality is encouraging to continue with our GLM-based DSR tool support idea, more research is required to improve the suggestion quality and to carefully design of the usage process, given the characteristics of GLM and the known effects AI can have on humans’ ways of working. The proposals of survey participants regarding design components, features and the usage process might be a good starting point. More broadly, future research should investigate the applicability of our approach to other fixed structures such as established canvasses (e.g., business model canvas) to develop GLM-based tool support in a structured way.

18.7 Limitations

There are some limitations to our study. Only one study (of Poser et al. (2022)) was included as a foundation for our survey, limiting representativeness. We decided against including more studies as the survey was already extensive. Including more studies would require participants to familiarize themselves with more, new research context(s) (incl. reading additional abstracts), increasing the burden to complete the survey. Indeed, we see

that most participants who did not complete the study stopped at the first page, which included the study details, strengthening us in the trade-off we made between completion (rate) and representativeness. Even with only one study included our sample size is small.

While we used tested prompt engineering techniques, we did not provide demonstration examples from other studies. A different presentation of the study contents towards the GLM or the utilization of other prompt engineering techniques might lead to different AI suggestions, potentially affecting perceived helpfulness. Additionally, including demonstration examples can improve suggestions quality for some tasks (Brown et al., 2020). In our study, we established a first baseline. Future studies, however, should explore appropriate configurations (prompt design, inference parameters) for the task. This might include conditioning the GLM to provide more unexpected or directive suggestions. In the study at hand, the GLM reproduced the structure that was entered, i.e., if the design requirement was phrased a certain way, the GLM would try to match this structure. With more elaborate prompt design, good practices for formulating design requirements (e.g., Chandra et al., 2015; Gregor et al., 2020) could be used.

18.8 Conclusion and outlook

Improving our understanding of how humans and AI can team up to solve complex problems is a new research challenge (Dellermann et al., 2019). For DSR, a particular approach to solving complex problems, we demonstrated the potential of support through AI, more specifically, through GLM. With our approach, we show that by making relevant study concepts available to a GLM in a structured way, helpful suggestions can be generated. In this way, facilitative DSR tool support can be complemented by an additional layer on content-level. While human collaboration remains essential for DSR projects, our results show that GLMs might be capable of contributing during periods of individual work.

Though the general perception of the AI suggestions was positive much more research is required. In order for the tool to be helpful, AI suggestions need to be of high quality. We therefore suggest to more qualitatively explore the reasons for (dis-)satisfaction with the suggestions and the underlying individual difference. To do so, the approach described in section 4.1 should be applied to other DSR studies, potentially extended by (some) suggestions for additional design components (section 5). Future research should explore the application to other, canvas-based problem solving tasks. Additionally, effects of prompt engineering (incl. demonstration examples, Zhu & Luo, 2022), parameter tuning, and an interplay of a hybrid approach using a generative and discriminative AI might be explored.

But even then, as pointed out by a participant, the sociotechnical usage process needs to be designed carefully. Challenges like overreliance or fatigue are known from other human-AI contexts (Buçinca et al., 2021; Seeber, Waizenegger, et al., 2020). The potential negative impact of external cues on creativity is known from design (Jansson & Smith, 1991).

Applicability of this existing design knowledge (e.g., interaction processes, interface design) for human-AI teams solving complex problems should be investigated.

To fully understand the effects of such tools it will be vital to have actual users develop actual concepts with such a tool. Do researchers adjust their concept due to GLM-based suggestions? How will the resulting DSR concepts compare to those without GLM-based tool support? The high level of interest among those who completed our survey, both to offer providing additional ideas and for using or testing a GLM-based DSR support tool, motivates us to continue exploring this path of human-AI collaboration.

Acknowledgement

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

18.9 References

- Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, 53(8), 18–28.
- Bedwell, W. L., Wildman, J. L., DiazGranados, D., Salazar, M., Kramer, W. S., & Salas, E. (2012). Collaboration at work: An integrative multilevel conceptualization. *Human Resource Management Review*, 22(2), 128–145. <https://doi.org/10.1016/j.hrmr.2011.11.007>
- Benke, I., Feine, J., Venable, J. R., & Maedche, A. (2020). On Implementing Ethical Principles in Design Science Research. *AIS Transactions on Human-Computer Interaction*, 12(4), 206–227. <https://doi.org/10.17705/1thci.00136>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*.
- Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–21. <https://doi.org/10.1145/3449287>
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C [Chiyuan]. (2023). Quantifying Memorization Across Neural Language Models. In *The International Conference on Learning Representations*.
- Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive Knowledge in IS Research: Conceptualizing Design Principles in Terms of Materiality, Action, and Boundary Conditions. In *48th Hawaii International Conference on System Sciences*, USA.

- Contell, J. P., Díaz, O., & Venable, J. R. (2017). DScaffolding: A Tool to Support Learning and Conducting Design Science Research. In A. Maedche, J. vom Brocke, & A. Hevner (Eds.), *Lecture Notes in Computer Science. Designing the Digital Transformation* (Vol. 10243, pp. 441–446). Springer International Publishing. https://doi.org/10.1007/978-3-319-59144-5_28
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition*. ACM.
- Durani, K., Eckhardt, A., & Kollmer, T. (2021). Towards Ethical Design Science Research. In *Austin, Texas*. https://aisel.aisnet.org/icis2021/adv_in_theories/adv_in_theories/3
- Fischer, A., Greiff, S., & Funke, J. (2012). The Process of Solving Complex Problems. *The Journal of Problem Solving*, 4(1). <https://doi.org/10.7771/1932-6246.1118>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Gau, M., Maedche, A., & vom Brocke, J. (2022). DSR Buddy: A Conversational Agent Supporting Design Science Research Activities. *Wirtschaftsinformatik 2022 Proceedings*, 35. https://aisel.aisnet.org/wi2022/prototype_track/prototype_track/35/
- Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing using Language Models. In F. ` Mueller, S. Greuter, R. A. Khot, P. Sweetser, & M. Obrist (Eds.), *Designing Interactive Systems Conference* (pp. 1002–1019). ACM. <https://doi.org/10.1145/3532106.3533533>
- Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards Designing Cooperative and Social Conversational Agents for Customer Service. In *38th International Conference on Information Systems*. <https://aisel.aisnet.org/icis2017/HCI/Presentations/1>
- Goldkuhl, G. (2004). Design Theories in Information Systems - A Need for Multi-Grounding. *JITTA : Journal of Information Technology Theory and Application*, 6(2), 59–72. <https://aisel.aisnet.org/jitta/vol6/iss2/7>
- Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, 37(2), 337–355. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Gregor, S., Kruse, L., & Seidel, S. (2020). Research Perspectives: The Anatomy of a Design Principle. *Journal of the Association for Information Systems*, 21, 1622–1652. <https://doi.org/10.17705/1jais.00649>
- Hevner, March, Park, & Ram (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), 75. <https://doi.org/10.2307/25148625>

- Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2). <https://aisel.aisnet.org/sjis/vol19/iss2/4>
- Hevner, A. R., Chatterjee, S., Gray, P., & Baldwin, C. Y. (2010). *Design research in information systems: Theory and practice. Integrated series in information systems: Vol. 22*. Springer.
- Jansson, D. G., & Smith, S. M. (1991). Design Fixation. *Design Studies*, 12(1), 3–11. [https://doi.org/10.1016/0142-694X\(91\)90003-F](https://doi.org/10.1016/0142-694X(91)90003-F)
- Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J. (2022). PromptMaker: Prompt-based Prototyping with Large Language Models. In *CHI '22: CHI Conference on Human Factors in Computing Systems* (pp. 1–8). ACM. <https://doi.org/10.1145/3491101.3503564>
- Kuechler, W., & Vaishnavi, V. (2012). A Framework for Theory Development in Design Science Research: Multiple Perspectives. *Journal of the Association for Information Systems*, 13(6), 395–423. <https://doi.org/10.17705/1jais.00300>
- Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.
- Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>
- Meth, H., Mueller, B., & Maedche, A. (2015). Designing a Requirement Mining System. *Journal of the Association for Information Systems*, 16(9), 799–837. <https://doi.org/10.17705/1jais.00408>
- Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 589–612).
- Morana, S., Scheid, M., Gau, M., Benke, I., vom Brocke, J., Fettke, P., & Maedche, A. (2018). Research Prototype: The Design Canvas in MyDesignProcess.com. In *13th International Conference on Design Science Research in Information Systems and Technology*.
- Morana, S., vom Brocke, J., Maedche, A., Seidel, S., Adam, M. T. P., Bub, U., Fettke, P., Gau, M., Herwix, A., Mullarkey, M. T., Nguyen, H. D., Sjöström, J., Toreini, P., Wessel, L., & Winter, R. (2018). Tool Support for Design Science Research—Towards a Software Ecosystem: A Report from a DESRIST 2017 Workshop. *Communications of the Association for Information Systems*, 43, 237–256. <https://doi.org/10.17705/1CAIS.04317>
- Myers, M. D., & Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information & Management*, 51(6), 801–809. <https://doi.org/10.1016/j.im.2014.01.002>

- Peffer, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Poser, M., Küstermann, G. C., Tavanapour, N., & Bittner, E. A. C. (2022). Design and Evaluation of a Conversational Agent for Facilitating Idea Generation in Organizational Innovation Processes. *Information Systems Frontiers*. Advance online publication. <https://doi.org/10.1007/s10796-022-10265-6>
- Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., & Sherson, J. (2021). Deskill, Upskill, and Reskill: a Case for Hybrid Intelligence. *Morals & Machines*, 1(2), 24–39. <https://doi.org/10.5771/2747-5174-2021-2-24>
- Rhys Cox, S., Wang, Y., Abdul, A., Weth, C. von der, & Y. Lim, B. (2021). Directed Diversity: Leveraging Language Embedding Distances for Collective Creativity in Crowd Ideation. In *CHI '21, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–35). Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445782>
- Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2). <https://doi.org/10.1016/j.im.2019.103174>
- Seeber, I., Waizenegger, L., Seidel, S., Morana, S., Benbasat, I., & Lowry, P. B. (2020). Collaborating with technology-based autonomous agents. *Internet Research*, 30(1), 1–18. <https://doi.org/10.1108/INTR-12-2019-0503>
- Sein, Henfridsson, Purao, Rossi, & Lindgren (2011). Action Design Research. *MIS Quarterly*, 35(1), 37. <https://doi.org/10.2307/23043488>
- Sio, U. N., Kotovsky, K., & Cagan, J. (2015). Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies*, 39, 70–99. <https://doi.org/10.1016/j.destud.2015.04.004>
- Stokel-Walker, C. (2023). Chatgpt listed as author on research papers: Many scientists disapprove. *Nature*, 613(7945), 620–621. <https://doi.org/10.1038/d41586-023-00107-z>
- Suh, M., Youngblom, E., Terry, M., & Cai, C. J. (2021). AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3411764.3445219>
- Venable, J. (2006). The Role of Theory and Theorising in Design Science Research. In *1st International Conference on Design Science in Information Systems and Technology*.

- vom Brocke, J., Fettke, P., Gau, M., Houy, C., & Morana, S. (2017). Tool-Support for Design Science Research: Design Principles and Instantiation. *SSRN Electronic Journal*. Advance online publication. <https://doi.org/10.2139/ssrn.2972803>
- vom Brocke, J., Winter, R., Hevner, A., & Maedche, A. (2020). Special Issue Editorial – Accumulation and Evolution of Design Knowledge in Design Science Research: A Journey Through Time and Space. *Journal of the Association for Information Systems*, 21(3), 520–544. <https://doi.org/10.17705/1jais.00611>
- Zhang, C [Chao], Yao, C., Liu, J., Zhou, Z., Zhang, W., Liu, L., Ying, F., Zhao, Y., & Wang, G. (2021). StoryDrawer: A Co-Creative Agent Supporting Children’s Storytelling through Collaborative Drawing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3411763.3451785>
- Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, 2, 1825–1834. <https://doi.org/10.1017/pds.2022.185>

19 Publication 10: Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy

Memmert, L., Borchers, M., Plückhahn, J., & Bittner, E. (2025). Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy. In *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, HI, USA. <https://hdl.handle.net/10125/109714>

Abstract

With the advancement of large language models (LLMs) such as ChatGPT, students increasingly employ LLMs during university courses to improve their coding skills. However, it is still unclear what using these systems means for student learning. In our study, we explore how students employ a LLM when solving a realistic coding challenge over the course of a semester in small groups. We analyze students' LLM chat log data through the framework of Bloom's taxonomy, a well-established education framework, to understand their problem-solving and learning behavior. We enrich our analysis with students' responses to weekly survey questions, contextualizing our findings with subjective experiences. We find that students primarily use LLMs for lower-order thinking skills like remembering, understanding, and applying procedural knowledge. With our study, we contribute to the growing literature on understanding the effects of working and learning with LLMs and offer practical suggestions for teachers and students.

Keywords: Coding, Hybrid Intelligence, ChatGPT, Bloom's Taxonomy, Education

19.1 Introduction

Coding education is an integral part of many computer science and information systems curricula. To effectively learn coding skills, students must apply theoretical skills and knowledge to real-world coding projects (Derus & Ali, 2012). Completing a project in a university course requires the integration of many skills, which can be overwhelming—particularly when working outside of the course sessions with no tutor present, students may be insecure (Nuutila et al., 2008).

LLMs like ChatGPT were shown to be capable of supporting learning and problem-solving (Dell'Acqua et al., 2023; Memmert & Bittner, 2024) in general and supporting (learning) coding activities in particular (Hartley et al., 2024). We suggest that using LLMs when working on a real-world coding project might help students complement human tutors' support. This aligns with recent reports of many students already using LLMs broadly (Al Shloul et al., 2024). However, it is unclear how students leverage LLMs and how LLMs support their learning and problem-solving journey. Thus, we seek to answer the two research questions:

Research question 1. How do students use a LLM for their coding projects?

Research question 2. How do students perceive the support of the LLM during their coding projects?

To answer these research questions, we offered students University-provided ChatGPT-Plus accounts to work on their projects (usage was not considered for grading) to ensure access to the ChatGPT chat log files. We analyzed these logs collected throughout two semesters of a university seminar in which students developed solutions for data-related coding challenges of a local company. As students primarily work on their projects outside of the in-presence seminar meetings, analyzing students' prompts offers unique insights into how students leverage the LLM to aid their problem-solving and learning journey. For analysis, we use a well-established, revised version of Bloom's Taxonomy as a deductive framework, allowing us to understand along which cognitive processes students employ the LLM. We enrich our analysis with responses to a weekly questionnaire in which students reflect on their experiences using a LLM. Due to the naturalistic data collection throughout the entire semester, there are limitations, e.g., students might have used other systems or deleted requests. Nonetheless, the collected data offers rich insights into students' LLM usage.

We find students to use ChatGPT primarily for low-order thinking skills like remembering and understanding procedures as well as the higher-order skill of creating. Although literature shows that LLMs can offer effective support when evaluating code (Popovici, 2024), we find students to rarely use it for code evaluation, which might be untapped potential.

With our work, we contribute to the literature on hybrid intelligence (Dellermann et al., 2019) characterizing the students' journey when working and learning with AI systems, as well as to literature on LLM use in academic contexts (Popovici, 2024). We offer practical suggestions for both teachers and students to more appropriately leverage the opportunities LLMs can offer in a learning context.

19.2 Background

19.2.1 Coding (education) in the age of large language models

Writing code can be perceived as challenging by students (Becker et al., 2023). While learning the theoretical foundation is key, applying it to real-world problems is an integral part of the learning journey (Derus & Ali, 2012), as students need to overcome many obstacles on their path to a solution. With recent advancements, LLMs can now support in educational contexts (Memmert et al., 2023; Park & Ahn, 2024) and students increasingly adopt them (Ibrahim et al., 2023). LLMs can help to tackle real-world coding challenges, e.g., by motivating students increasing their self-efficacy (Deriba et al., 2023), by explaining or simplifying complex code (Becker et al., 2023; Chen et al., 2023; Deriba et al., 2023), by providing code (examples) to complete a specific task (Becker et al., 2023; Deriba et al., 2023), by reviewing code and debugging (Becker et al., 2023; Deriba et al., 2023; Lififiton et al., 2023), or by explaining concepts (Deriba et al., 2023). Students felt using LLMs is helpful for learning programming (Groothuijsen et al., 2024)—LLMs can enhance task performance even for experienced programmers (Mehler & Krautter, 2024).

Therefore, in the future, educators will need to consider if and how they embed LLMs into their courses and how they educate students on using LLMs (Becker et al., 2023; Groothuisen et al., 2024). In this study, we explore how students currently use LLM in real-world coding challenges to uncover both frequent areas of using and not using LLMs. This could serve as a foundation for future research to explore the appropriateness of using LLMs for different aspects in a challenge-based learning context. A frequently used framework for structuring the discussion around learning is Bloom's Taxonomy (Bloom et al., 1956), which we explain in more detail below.

19.2.2 Bloom's (revised) taxonomy

Bloom's taxonomy (Bloom et al., 1956) is a well-established taxonomy from education literature that was originally introduced in 1956 to aid the communication around learning objectives and structure learning goals and assessments. In 2001, the taxonomy was revised (Anderson et al., 2001) and is hereafter referred to as revised taxonomy. The revised taxonomy breaks down learning into six layers with increasingly complex cognitive processes, moving from lower-order to higher-order thinking skills. Krathwohl (2002) described the six layers as follows:

1. *Remember* refers to retrieving knowledge from memory
2. *Understand*, though sometimes used more broadly, refers to determining meaning
3. *Apply* refers to carrying out a procedure
4. *Analyze* refers to breaking down material into its parts to detect their interrelations
5. *Evaluate* refers to criteria-based judging
6. *Create* refers to developing something novel

The layers are hierarchical, i.e., in general, higher layers are assumed to be more complex, though there can be overlaps (Krathwohl, 2002). While the hierarchical nature was discussed critically in literature, empirical evidence suggests a progression from lower- to higher-order thinking skills (Athanasios et al., 2003). Besides the six cognitive process dimensions, the taxonomy includes a breakdown of four knowledge types. These are essential to enhance the application and effectiveness of Bloom's taxonomy in education settings and are explained by Krathwohl (2002) as follows:

- A. *Factual* refers to the basic elements of a topic
- B. *Conceptual* refers to the relationships between the elements and the larger structure as a whole
- C. *Procedural* refers to the way of doing something, such as methods, techniques, and approaches
- D. *Metacognitive* refers to breaking down material into its parts to detect their interrelations

These two dimensions of cognitive process and knowledge type together create a table, that can be used to classify different learning-related aspects, such as goals and objectives

(Kiesler, 2020; Krathwohl, 2002). Such a table can also be used for “gap spotting”, i.e., identifying underrepresented combinations of cognitive processes and knowledge types in a learning journey (Krathwohl, 2002).

19.2.3 Bloom’s taxonomy and (coding) learning support tools

It was questioned if using Bloom’s (revised) taxonomy is appropriate for computer science (Fuller et al., 2007; Johnson & Fuller, 2006), e.g., given that learning to code seems inherently like students need to *apply* and *create* (code). In the light of our study, in which we explore how students use LLMs to solve real-world coding challenges, this question might seem particularly relevant, with students seeking both to solve a problem and to advance their coding skills. However, it was noted that while “problem-solving” is a term frequently used by educators in the context of learning, it does not strictly relate to one of the cognitive processes layers but is related to different layers depending on context (Krathwohl, 2002). Thus, even when students solve problems, the cognitive processes they engage in at each stage can relate to different taxonomy layers. Additionally, while the original idea of Bloom’s taxonomy was to use it for structuring the communication around learning goals, over the last decades it was used in numerous ways to analyze computer science and coding-related learning progress across different stages (Ullah et al., 2020) even across longer projects (Alaghbary, 2021), to classify coding exercises and enhance teaching practices and exams (Dorodchi et al., 2017; Masapanta-Carrión & Velázquez-Iturbide, 2019; Selby, 2015), for classifying software including coding-related tools (Xu et al., 2009), and to classify problem-solving behavior in computer science (Parham et al., 2009). This includes research on if and how students move between the different layers, with Athanassiou et al. (2003) finding students to achieve higher-order thinking skills later. Parham et al. (2009) find that most successful students frequently switch between levels, contrary to the taxonomy’s hierarchical nature and prior literature.

We seek to analyze the students’ journey of learning to solve a data challenge. Bloom’s revised taxonomy was widely used as a framework to make sense of learning-related aspects with respect to coding. We, thus, suggest using it as a foundation here. While the taxonomy is widely used to analyze students’ journeys, when students interact with LLMs on a certain level, this does not imply that students actually acquired the corresponding knowledge, but rather, the analysis uncovers which cognitive processes students are engaged in with respect to which knowledge type.

19.3 Method

19.3.1 Data collection

We conducted a study in a university course repeated across two semesters with different students. We had 31 students, of which 15 participated in the first and 16 in the second

semester. The students worked in groups of five to six students on data-related coding projects for which a major public transportation company provided real-world use cases, e.g., problems and data. The group projects spanned the entire semester and resulted in an executable "prototype," demonstrated to the company during a final presentation. All challenges address public transport tasks and cover the automated evaluation of sustainability reports, the analysis of websites for stakeholder searches, the analysis of weather events for disruptions to public transport, the prediction of delays in subway operations, and the frequency of vehicle use. Even if the challenges pursue different goals, they are comparable. To develop solutions, e.g., the prototype, the students used data from the company, freely available data, and new or existing AI models.

Throughout the semesters, each team had access to a LLM account (ChatGPT 4 Plus) to assist throughout the project. Students were free to use or not use the LLM (usage was neither mandatory nor considered for grading). To understand the actual work with the LLM, we exported the chat log data of the groups that used the LLM. For the perceived experience, we conducted a weekly survey with open-ended questions (survey data), asking how students integrated ChatGPT into their work in the past week and how working with the system supported their learning. Collecting data on a weekly basis ensures the experiences are still fresh and remembered by the students. Combining both data sources (chat log and survey data) allows us to contrast actual work behavior and perceived experience.

19.3.2 Data analysis

In our study, we seek to understand how students employ LLMs during their learning journey when working on a real-world coding challenge. For the analysis, we exported all the students' conversations with the LLM. As discussed, Bloom's (revised) taxonomy is a well-established framework for making sense of the different levels of learning. Therefore, we performed a qualitative content analysis (Mayring, 2014) on the LLM chat logs following a deductive coding approach, using a revised version of Bloom's taxonomy as a deductive frame. We performed the analysis based on the prompt-level instead of conversation-level, as prior research showed that even single exercises can affect multiple cognitive processes (Masapanta-Carrión & Velázquez-Iturbide, 2019). To understand these processes, our analysis focused on the student's prompts. We included the LLM output only to provide additional context if required.

We had a trained assistant classify all student prompts according to the levels and knowledge type of Bloom's revised taxonomy. The coding schema was developed based on prior literature (Kiesler, 2020; Nikolić & Dabić, 2016; Thompson et al., 2008) with the conceptualizations of the cognitive domain levels as described in Table 1. Similarly to Kiesler (2020), besides the cognitive domain levels of Bloom's taxonomy, we applied the knowledge types as a secondary dimension to further classify the prompts, as shown in Table 2.

Table 1. Levels and explanations

Level	Explanation
Remember	Remembering foundational terminology and commands of programming language and foundational functions of the integrated development environment (IDE)
Understand	Understanding code and understanding code writing as well as the functionality of the programming language and ChatGPT
Apply	Reaching certain programming goals by applying known (or to be known) concepts to new data, i.e., it is not a creative but an operational approach
Analyze	Analysis of code and error messages for supporting debugging/ error fixing
Evaluate	Evaluation of inputs through queries, potential interpretations, or assessments as well, as requests for suggestions for improvements
Create	Generating something novel, such as code, text, or ideas, with a focus on it being a creative activity

Table 2. Knowledge types and explanations

Knowledge	Explanation
Factual	Concrete knowledge of fundamental information and terms of programming or general work with an IDE
Conceptual	Deeper knowledge regarding the connection of fundamental information and concepts of programming, as well as the characteristics of algorithms and the understanding of their structures
Procedural	Application-oriented knowledge regarding the solution of problems using algorithms. This includes the understanding, implementation, and evaluation of algorithms
Metacognitive	Abstract knowledge of one's own working methods and cognitive abilities. This includes procedural planning when writing algorithms or executing other project work, as well as self-monitoring and self-assessment

All prompts that referred to requests not related to the project (e.g., working on tasks from different seminars, or for private purposes), non-learning-related requests, repetitions of the same request (e.g., due to lack of LLM response), requests with images or files as part of the prompt, were excluded.

To get a better understanding of LLM usage, we report descriptive statistics regarding prompting behavior and offer qualitative insights into how the LLM is used across different levels. We perform a time-based analysis by calculating the time difference of each prompting time to the first prompt request made during this semester across groups as the students did not directly start using the LLM in the first weeks of the seminar in which a general introduction to the seminar was given by the instructors. We then enrich our data by analyzing the responses to the open-ended weekly survey questions to better understand the perception of the students working and learning with a LLM.

19.4 Results

In total, we analyzed data from 31 students in 6 groups and 637 prompts. The prompts were part of 142 conversations, with an average number of prompts per conversation of 4.49. We first analyzed the chat logs regarding Bloom's taxonomy. Figure 1 offers an overview of the frequency of LLM usage in relation to the levels of Bloom's revised taxonomy. Most of the prompts fall into the lower-order thinking skills from *remember*, *understand*, and *apply*. Higher-order thinking skills are less frequently used, except for *create*.

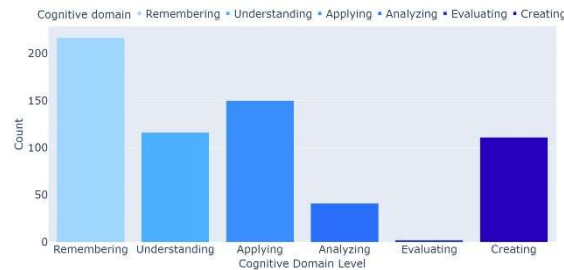


Figure 1. Prompt frequency regarding cognitive domain levels in Bloom's revised taxonomy

For *remember*, prompts included questions regarding certain commands, e.g., opening files, reaching specific windows/views, or functionality of buttons within the applications.

To *understand*, students prompted the LLM about certain code snippets, about how ChatGPT works and how to work with it, understanding certain commands, programming concepts (e.g., classes, methods), or entire programming languages, or commenting code.

In prompts regarding *apply*, students inquired in a more detailed manner on how to achieve certain coding goals, in being supported in data processing (e.g., creating tables, but not creating the entire code), and more generally, applying known concepts to new data (e.g., rewriting code or formulating regular expressions).

Analyze included understanding error messages and the related code. Students almost never used ChatGPT to evaluate. In the two cases identified, users asked for feedback on the code. For *create*, students had ChatGPT create code, and parts of the code generated text or ideas, i.e., something new. In several cases, the LLM created long passages of non-trivial code. Students asked follow-up questions regarding the code provided by the LLM, e.g., to understand the contents of variables, with one student asking: “is matches a boolean?”

While we did not monitor the usage of the LLM's output, chat logs seem to indicate students have copied code into their projects. They then ask for modifications of the code or copy error messages they receive during subsequent code execution, e.g., with one student explaining to the LLM: „I have implemented your suggestions, and now I get those errors after I enter something in a URL in the text field and press the button. Here is the error message: [long error message]“. Based on those follow-up questions and the error messages, it seems students used the LLM code without fully understanding it.

Regarding the literature, we expected some progression when working on the tasks, from lower, order thinking skills to higher-order thinking skills (Athanassiou et al., 2003). When plotting the prompts classified against the cognitive domain levels of Bloom's revised taxonomy over time (see Figure 2), we do not find such a trend to occur.

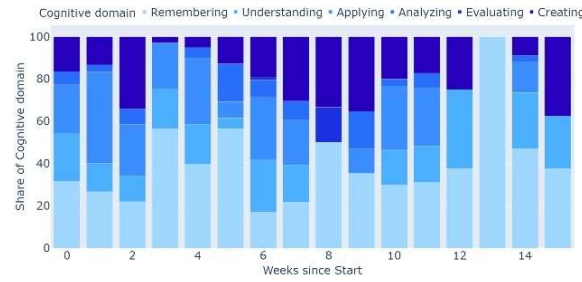


Figure 2. Share of prompts classified according to cognitive domain levels of Bloom's revised taxonomy over time

As a second dimension, in addition to the cognitive domain levels of Bloom's revised taxonomy, the prompts were classified according to their knowledge type. Figure 3 offers an overview of the respective frequencies.

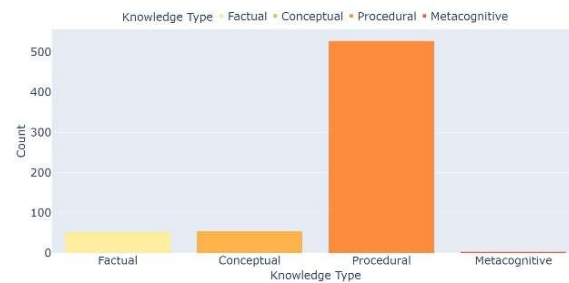


Figure 3. Prompt frequency regarding knowledge levels

Students mostly wrote prompts regarding *procedural* knowledge, i.e., to understand how to do something, such as writing a piece of code to achieve something. Sometimes, students prompted the LLM to understand certain concepts (*conceptual* knowledge), such as software design patterns or code execution in threads. Additionally, the LLM was used for *factual* knowledge, e.g., to understand which parameters a library function accepts. Rarely were the prompts concerned with the metacognitive level, e.g., when students asked for a plan to approach the project in general or, e.g., to create a plan for refactoring certain code. When plotting the levels of knowledge over time, we find procedural knowledge to be most important throughout the semester. Only at the end of the semester were procedural and conceptual knowledge reflected similarly.

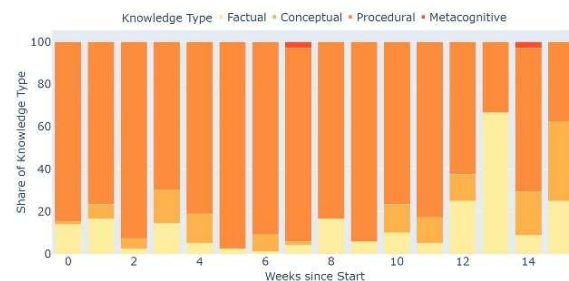


Figure 4. Share of prompts classified according to knowledge types over time

When combining both the cognitive domain levels of Bloom's revised taxonomy and the knowledge type axes (Figure 5), we find that students most frequently sought to apply a

procedure (applying-procedural: 150), remember how to do something (remembering-procedural: 149), or create new procedures (creating-procedural: 108). Outside of the procedural knowledge type, we find remembering facts to be most relevant (remembering-factual: 52). We find examples for most of the level-knowledge type combinations identified by Kiesler (2020) except *conceptual-analyzing* and *metacognitive-evaluating*.

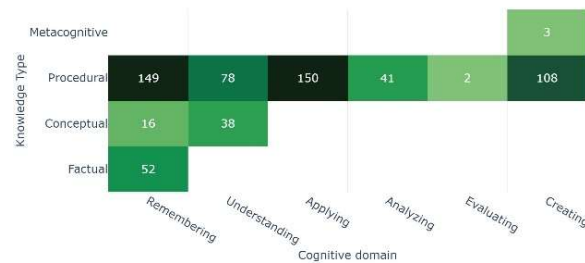


Figure 5. Heat map for cognitive domain levels across knowledge types

Analyzing the responses to the weekly survey questions offers more context to the findings from the chat logs, such as the high number of prompts related to factual and *procedural remembering* and *understanding*. Students reported having used the LLM instead of “tedious[ly] searching through library documentations or hoping somebody asked something similar on Stackoverflow” and explained, “it is way easier to learn with chatgpt and analyse and update the code instead of watching tutorials on YouTube for hours.” Using the LLM instead of Google, YouTube, Stackoverflow, asking the teacher, or reading the documentation was stated to be “much easier” and “much quicker than reading documentation.”

From a more *conceptual* perspective, students seem to have perceived the LLM to, in general, get a better understanding of their topic, explaining, “I would definitely recommend [ChatGPT] for getting an idea of what to investigate more into, as well as for getting some insight in topics you are less familiar with (generally better suited than simply googling unfamiliar topics)” and that ChatGPT “gives good topic overviews.” It also helped them to develop ideas, with a student stating ChatGPT to be “pretty good to verify ideas you have. Just to make sure that the idea is not completely in the wrong direction” and to approach coding in new ways, stating, “[ChatGPT] was helpful by showing me how to use new ways to solve my coding problem” and “[ChatGPT] is good to learn something about coding or to get an idea of how to implement things.”

Beyond using the LLM for lower-order thinking skills, working with it seems to have changed the way of learning. Many students explained that they used the LLM to get a first starting point, on which they then improved upon, explaining, “I let [ChatGPT] write basic scripts so the basic building block of everything. Then I changed some functions or wrote new ones. At the end I gave [ChatGPT] the whole script for feedback.” Several students explained they believed ChatGPT supports learning, stating “that it can be very helpful for beginners using a programming language.” They highlighted that ChatGPT helped them to focus on learning, explaining that “[ChatGPT] helped [them] to learn more about build-

ing/using neural networks by making [them] spend less time on bug fixing” and that learning was supported by the LLM producing “real code examples for [their] problems.” According to the students, ChatGPT can support learning, particularly when used with caution, with one student stating: “Great for learning. But not give its answers too much authority. Seeing ChatGPT like a learning partner that has good enough knowledge but makes critical mistakes and has knowledge gaps”.

However, other workflows seem less optimal from a learning perspective, with a student displaying more of a trial-and-error approach as opposed to active reflection when explaining: “the first response won’t give you the right answer. Send the error message / what went wrong and chatgpt will send you a better solution. Until it’s correct (hopefully).” When reflecting on their usage, students shared some concerns, stating that “asking for code implementations was mostly distracting,” that using ChatGPT is “time efficient [...] but that to learn to code using it is a bit of a cheat”, and “in most cases either wrong, not existing or very complicated results were created” by ChatGPT. It was even questioned if ChatGPT helps at all with learning, with a student stating, “For coding it doesn’t help learning, because it takes work from you both debugging and writing.”

19.5 Discussion

19.5.1 General discussion

We investigated the LLM usage of student groups working on a real-world coding challenge through the lens of (a revised version of) Bloom’s taxonomy. Below, we discuss the (non-)observed behavior and challenges in students using a LLM for learning to code and reflect on using Bloom’s taxonomy as a lens for analysis.

(Non-)Observed behavior. We find students mostly make requests related to lower-order thinking skills, i.e., the three top-most request types are *remember*, *understand*, and *apply*, all of which are lower-order thinking skills. This suggests that students partially use the LLM to replace using search engines, forums (e.g., stackoverflow.com), or documentation, which was also reflected in the survey responses.

From the higher-order thinking skills, the top-most level of *creating* is the most frequently used. This is not surprising, as solving real world coding challenges involves developing new code. Using the LLM helps students by generating code as a starting point, which they can then adapt, benefitting novices of not having to start from scratch (Becker et al., 2023; Jonsson & Tholander, 2022). Many of the students seemed to have an idea of what they wanted to achieve and used the LLM to translate their thoughts into code, which aligns with findings that LLMs can prevent students from getting “stuck” in trying to express their thoughts in the programming language (Jonsson & Tholander, 2022).

However, this approach can lead to problems. In several cases, after using the LLM to create code, students made follow-up requests with error messages regarding the code.

Though we did not track how students handled the LLM's output, based on the follow-up request, the students seem to have copied the code into their project, perhaps without (fully) understanding and checking the code (at least they did not seem to have corrected the error themselves). This could hint at the known problem of overreliance when students work with LLMs (Becker et al., 2023; Deriba et al., 2023). This was also reflected in the diary responses regarding the students' approach towards fixing code errors, such as "send the error message [...] until its correct (*hopefully*)" (emphasize added).

Another higher-order thinking skill is *evaluate*. As discussed by Krathwohl (2002), Bloom's taxonomy can be used for gap spotting. In our study, students rarely used the LLM to evaluate their code, despite one student explaining a workflow using the LLM to get feedback on their code. This rare usage of evaluation could hint at a gap and perhaps an untapped potential, given that prior research showed that LLMs like ChatGPT are good at offering code reviews (Popovici, 2024).

Similarly, only a few prompts related to metacognitive knowledge. This could suggest that students mostly performed the planning and strategy development for working on the tasks or approached the tasks without prior strategies. Overall, LLMs seem to have been predominantly used for execution and implementation instead of planning, even though Jonsson and Tholander (2022) pointed out that LLMs can support structuring. Some survey responses confirmed that students found it helpful to get the LLM's "perspective" on approaching parts of the task.

Challenges when using LLMs for learning to code. Tools can be classified according to Bloom's taxonomy in which aspects of learning they are capable of supporting (Xu et al., 2009). We find students used ChatGPT across all levels, though to varying degrees, which comes with several challenges.

Even if students use the LLM across levels, it is unclear how well LLMs are suited for different purposes. There are some system-based challenges to overcome, such as hallucination (Susarla et al., 2023), which could result in the LLMs generating code relying on supposedly existing library functions, which do not exist, particularly problematic for requests on remembering factual knowledge. Thus, teachers should make students aware of this problem and explain how to effectively work in such cases, e.g., by checking the library documentation in question. In the diary responses, we generally saw a high awareness among the students about this issue.

Besides problems of the system itself, challenges can also arise in the interaction. Previous research pointed out that prompting LLMs effectively can be a challenge for novices more generally (Zamfirescu-Pereira et al., 2023) and in the context of producing code specifically (Jonsson & Tholander, 2022), and that prompts significantly affect generated code quality (Becker et al., 2023). The chat logs contained repeated and slightly adjusted prompts and requests of the students towards the LLM to support them in finding better or aggregating existing prompts. One of the students even responded to the LLM, "I think you

misunderstood me. [...]”. Thus, effective prompting might be a key skill for students to produce code.

However, even when students learn to prompt, it is unclear if learning will be supported optimally. In the chat logs, we found several instances where students asked about lower-order thinking skill levels, but the LLM provided a code implementation. While this might help to produce results more quickly, it could reduce students’ learning as they receive the solution directly. A more targeted approach might be to develop a system that automatically detects the level of the student question (see, e.g., Reddy & Subair, 2017) and offers an answer aligned to the question’s level, leaving the problem-solving to the student.

Bloom’s Taxonomy as a Lens. The classification along the cognitive levels of Bloom’s taxonomy offers great insights into LLM usage, and we found Kiesler’s (2020) approach particularly helpful in further differentiating the prompts according to knowledge levels. Particularly for the level of remembering, the prompts had different structures, e.g., regarding requests for facts and requests on procedures (“how to”-requests). We found no conflict with the combinations of levels and knowledge types identified by Kiesler (2020). We found examples for most suggested combinations, generally confirming them (Figure 5). We did not find examples for *analyzing-conceptual* and *evaluating-metacognitive*. The lack of examples for the knowledge type metacognitive might explain the lack of an example for the latter combination.

Students are expected to progressively achieve higher-order thinking skills throughout the semester (Athanassiou et al., 2003). Contrary to this expectation, we did not find a large difference in LLM usage according to the taxonomy’s levels. Perhaps, as pointed out by Fuller et al. (2007), coding is already applied, or maybe because effective learning uses alteration between levels instead of a linear progression (Parham et al., 2009). A different explanation might be that there are many skills to master in a coding project. Thus, while students might have reached a degree of proficiency in one aspect (e.g., data ingestion), they might still need to learn a lot in other areas (e.g., data visualization or communication), which results in different requests at different levels towards the LLM.

19.5.2 Contributions

With our work, we contribute to the growing body of knowledge on how LLMs can benefit students in educational settings in general (Dwivedi et al., 2023; Ooi et al., 2023) and learning to code (Popovici, 2024). We offer orientation on how a conversational LLM like ChatGPT is used by students with limited coding experience during a realistic coding project across several months and how the students perceive working with the system. Our descriptive approach to understanding how LLMs are used serves as a foundation for more prescriptive knowledge, such as recommendations for educators or students on using LLMs for coding projects, assignments, or developing tools to support learning how to code.

Besides the body of knowledge on hybrid intelligence (in educational settings), we also contribute to the literature on using Bloom's revised taxonomy. We show that the taxonomy can be used to make sense of students' learning behavior during semester-long coding projects and to classify coding support by tools, as discussed, e.g., by Xu et al. (2009).

19.5.3 Implications

Our study has implications for teaching coding:

Encourage exploring LLMs. Overall, we find students to enjoy working with LLMs. Given recent findings regarding productivity enhancement through LLMs when coding, we strongly encourage including coding with LLMs in teaching coding in universities to prepare students to use LLMs sustainably in their jobs. Based on our findings, educators and students should consider several points described below.

Encourage using LLMs for code reviews. ChatGPT can offer good code reviews (Popovici, 2024), but according to our data, students rarely use ChatGPT to get feedback on their code. Thus, when teaching a coding course, instructors should encourage students to use ChatGPT to get feedback on their own code, and to identify design flaws as well as more trivial aspects like naming elements or commenting.

Encourage engagement with the code. Studies show that the code generated by ChatGPT is not necessarily readable for novices (Popovici, 2024). Thus, while students should leverage LLMs, we urge them to critically engage with the code to understand it and learn from it, e.g., by asking questions about the code. Several students followed this logic asking follow-up questions instead of merely copying the code. However, others seem to have simply copied the incorrect code, running into downstream errors.

Stress the importance of understanding the basics. Our findings indicate that students often used LLMs for lower-order thinking skills, replacing looking up certain functionality in the documentation. We suggest educators reiterate the value of reading and understanding code documentation. From a learning perspective, this is relevant due to the hierarchical nature of the processes, with higher-level cognitive processes necessitating lower-level processes, i.e., students require foundations to achieve higher-level goals (e.g., to make critical judgments of LLM outputs). From a practical perspective, this is also relevant, e.g., when working in organizations with a custom code base for which documentation is inaccessible to the LLM.

19.5.4 Limitations and outlook

There are several limitations to our study. When investigating how students worked with LLMs on their challenges, we only considered one LLM (ChatGPT). There are other (conversational) LLMs available, and even LLMs specifically for coding support, such as GitHub Copilot. However, ChatGPT is the most rapidly adopted technology and is used by students for coding assistance (Popovici, 2024), given its low entry barrier, which is why we selected it for our study. We do, however, encourage exploring other (integrated) LLMs

for coding, as some challenges such as efficient code provision will be addressed by such tools.

We decided to follow a naturalistic data collection throughout the semester to understand students' LLM usage for the projects. While this offers rich data, it comes with limitations. We offered students ChatGPT Plus access for free to incentivize participation, though it was neither mandatory nor considered for grading. However, students might have used other tools and could have deleted requests.

Bloom's taxonomy is widely used to develop new courses and classify existing content like teaching material, exams, or student questions. Even though widely used, classification according to the taxonomy is qualitative research and, thus, can be subjective. To limit this, a classification schema was developed based on literature (Kiesler, 2020; Nikolić & Dabić, 2016; Thompson et al., 2008), but a degree of ambiguity remains, particularly for longer prompts.

Future research might want to explore whether LLM prompts could be classified automatically. Nikolić and Dabić (2016) offer a list of verbs for each level of Bloom's taxonomy. This list could be used to build a classifier to code larger amounts of data (comparable to the tool developed by Reddy & Subair, 2017). Such an automatic classification could make the coding process more objective and enables a tool dedicated to supporting *learning* when coding.

We found that in many cases, the student asked on a specific level, i.e., to conceptually understand a coding aspect. However, the LLM not only provided the conceptual explanation but directly included the code for implementation. This proactivity, while perhaps helpful in a working setting, could hinder learning because it does not force students to engage with conceptual explanations and apply the knowledge to create code themselves. Automatic classification of the prompts according to Bloom's levels could help to develop a tool that only responds on the level on which it was asked. Future research might explore how such an approach could be operationalized, e.g., using LLMs for classification or prompt design for targeted responses.

As discussed, our study does not measure learning progress but analyzes for which types of cognitive processes and types of knowledge students turn to LLMs. Our time-based analysis did not show a clear progression from lower-order to higher-order thinking skills throughout the semester, as suggested by prior research (Athanassiou et al., 2003). Instead, we find students use several cognitive domains throughout the semester, a behavior more strongly aligned with the frequent-switching behavior suggested by Parham et al. (2009). Future research should explore if and how LLMs can help students reach those higher-order thinking skills more seamlessly. An additional and perhaps more important question is if achieving these higher-order thinking skills happens in a sustainable way, i.e., if students actually learn to code better in the long term instead of merely superficially editing code.

19.6 Conclusion

Students will use LLMs for coding in their future jobs. Universities should use the opportunity to shape such usage and support students' learning journey when working with LLMs. We show that Bloom's taxonomy (revised) is a helpful frame for understanding working and learning behavior. We make several suggestions for improving how instructors and students can use LLMs in educational settings and offer ideas on how to proceed with exploring LLM-supported teaching and learning of coding activities.

Acknowledgment

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

19.7 References

- Al Shloul, T., Mazhar, T., Abbas, Q., Iqbal, M., Ghadi, Y. Y., Shahzad, T., Mallek, F., & Hamam, H. (2024). Role of activity-based learning and ChatGPT on students' performance in education. *Computers and Education: Artificial Intelligence*, 6, 100219.
- Alaghbary, G. S. (2021). Integrating technology with Bloom's revised taxonomy: Web 2.0-enabled learning designs for online learning. *Asian EFL Journal*, 28(1), 10–37.
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.
- Athanassiou, N., McNett, J. M., & Harvey, C. (2003). Critical Thinking in the Management Classroom: Bloom's Taxonomy as a Learning Tool. *Journal of Management Education*, 27(5), 533–555. <https://doi.org/10.1177/1052562903252515>
- Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., & Santos, E. A. (2023). Programming Is Hard - Or at Least It Used to Be. In M. Doyle, B. Stephenson, B. Dorn, L.-K. Soh, & L. Battestilli (Eds.), *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (pp. 500–506). ACM. <https://doi.org/10.1145/3545945.3569759>
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R., & others. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman NY.
- Chen, E., Huang, R., Chen, H.-S., Tseng, Y.-H., & Li, L.-Y. (2023). GPTutor: A ChatGPT-Powered Programming Tool for Code Explanation. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, & O. C. Santos (Eds.), *Communications in Computer and Information Science. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky* (Vol. 1831, pp. 321–327). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36336-8_50

- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>
- Deriba, F. G., Sanusi, I. T., & Sunday, A. O. (2023). Enhancing Computer Programming Education using ChatGPT- A Mini Review. In A. Mühling & I. Jormanainen (Eds.), *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research* (pp. 1–2). ACM. <https://doi.org/10.1145/3631802.3631848>
- Derus, S. R. M., & Ali, A. M. (2012). Difficulties in learning programming: Views of students. In *1st International Conference on Current Issues in Education*.
- Dorodchi, M., Dehbozorgi, N., & Frevert, T. K. (2017). "I wish I could rank my exam's challenge level!": An algorithm of Bloom's taxonomy in teaching CS1. In *2017 IEEE Frontiers in Education Conference (FIE)*.
- Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71. <https://doi.org/10.1016/j.ijinfomgt.2023.102642>
- Fuller, U., Johnson, C. G., Ahoniemi, T., Cukierman, D., Hernán-Losada, I., Jackova, J., Lahtinen, E., Lewis, T. L., Thompson, D. M., Riedesel, C., & Thompson, E. (2007). Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin*, 39(4), 152–170. <https://doi.org/10.1145/1345375.1345438>
- Groothuisen, S., van den Beemt, A., Remmers, J. C., & van Leeuwen, L. W. (2024). AI chatbots in programming education: students' use in a scientific computing course and consequences for learning. *Computers and Education: Artificial Intelligence*, 100290. <https://doi.org/10.1016/j.caeai.2024.100290>
- Hartley, K., Hayak, M., & Ko, U. H. (2024). Artificial Intelligence Supporting Independent Student Learning: An Evaluative Case Study of ChatGPT and Learning to Code. *Education Sciences*, 14(2), 120. <https://doi.org/10.3390/educsci14020120>
- Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., Baghdadi, R., Bélanger, J. J., Beretta, E., Celik, K., Chaqfeh, M., Daqaq, M. F., Bernoussi, Z. E., Fournie, D., Soto, B. G. de, Gan-

- dolfi, A., . . . Zaki, Y. (2023). Author Correction: Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-43998-8>
- Johnson, C. G., & Fuller, U. (2006). Is Bloom's taxonomy appropriate for computer science? In A. Berglund (Ed.), *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006* (pp. 120–123). ACM. <https://doi.org/10.1145/1315803.1315825>
- Jonsson, M., & Tholander, J. (2022). Cracking the code: Co-coding with AI in creative programming education. In *Creativity and Cognition* (pp. 5–14). ACM. <https://doi.org/10.1145/3527927.3532801>
- Kiesler, N. (2020). Towards a Competence Model for the Novice Programmer Using Bloom's Revised Taxonomy - An Empirical Approach. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 459–465). ACM. <https://doi.org/10.1145/3341525.3387419>
- Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice*, 41(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
- Liffiton, M., Sheese, B., Savelka, J., & Denny, P. (2023). *CodeHelp: Using Large Language Models with Guardrails for Scalable Support in Programming Classes*.
- Masapanta-Carrión, S., & Velázquez-Iturbide, J. Á. (2019). Evaluating Instructors' Classification of Programming Exercises Using the Revised Bloom's Taxonomy. In B. Scharlau, R. McDermott, A. Pears, & M. Sabin (Eds.), *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 541–547). ACM. <https://doi.org/10.1145/3304221.3319748>
- Mayring, P. (2014). *Qualitative Content Analysis - Theoretical Foundation, Basic Procedures and Software Solution*.
- Mehler, M., & Krautter, K. (2024). Productivity vs. Purpose: Generative AI Enhances Task Performance but Reduces Meaningfulness in Programming. In *ECIS 2024 Proceedings*. https://aisel.aisnet.org/ecis2024/track05_fow/track05_fow/7
- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., Borchers, M., Plückhahn, J., & Bittner, E. (2025). Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy. In *Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)*, Waikoloa Village, HI, USA. <https://hdl.handle.net/10125/109714>
- Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. *Journal of Information Systems Education*, 34(3), 279–292. <https://aisel.aisnet.org/jise/vol34/iss3/3>

- Nikolić, M., & Dabić, T. (2016). The Bloom's Taxonomy Revisited in the Context of Online Tools. In *Proceedings of the International Scientific Conference - Sinteza 2016* (pp. 315–320). Singidunum University. <https://doi.org/10.15308/Sinteza-2016-315-320>
- Nuutila, E., Törmä, S., Kinnunen, P., & Malmi, L. (2008). Learning Programming with the PBL Method — Experiences on PBL Cases and Tutoring. In J. Bennedsen, M. E. Caspersen, & M. Kölling (Eds.), *Lecture Notes in Computer Science. Reflections on the Teaching of Programming* (Vol. 4821, pp. 47–67). https://doi.org/10.1007/978-3-540-77934-6_5
- Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., . . . Wong, L.-W. (2023). The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. *Journal of Computer Information Systems*. Advance online publication. <https://doi.org/10.1080/08874417.2023.2261010>
- Parham, J., Chinn, D., & Stevenson, D. E. (2009). Using Bloom's taxonomy to code verbal protocols of students solving a data structure problem. In J. D. McGregor (Ed.), *Proceedings of the 47th Annual Southeast Regional Conference* (pp. 1–6). ACM. <https://doi.org/10.1145/1566445.1566499>
- Park, H., & Ahn, D. (2024). The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *CHI '24 Proceedings*.
- Popovici, M.-D. (2024). ChatGPT in the Classroom. Exploring Its Potential and Limitations in a Functional Programming Course. *International Journal of Human-Computer Interaction*, 40(22), 7743–7754. <https://doi.org/10.1080/10447318.2023.2269006>
- Reddy, D., & Subair, R. (2017). Automated tool for Bloom's Taxonomy. *Technology*, 8(7), 544–555. <http://www.iaeme.com/ijciyet/issues.asp?JType=IJCIET&VType=8&IType=7>
- Selby, C. C. (2015). Relationships: computational thinking, pedagogy of programming, and Bloom's Taxonomy. In J. Gal-Ezer, S. Sentance, & J. Vahrenhold (Eds.), *Proceedings of the Workshop in Primary and Secondary Computing Education* (pp. 80–87). ACM. <https://doi.org/10.1145/2818314.2818315>
- Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, 34(2), 399–408. <https://doi.org/10.1287/isre.2023.ed.v34.n2>
- Thompson, E., Luxton-Reilly, A., Whalley, J. L., Hu, M., & Robbins, P. (2008). Bloom's taxonomy for CS assessment. In *Proceedings of the tenth conference on Australasian computing education*.

- Ullah, Z., Lajis, A., Jamjoom, M., Altalhi, A., & Saleem, F. (2020). Bloom's taxonomy: A beneficial tool for learning and assessing students' competency levels in computer programming using empirical analysis. *Computer Applications in Engineering Education*, 28(6), 1628–1640. <https://doi.org/10.1002/cae.22339>
- Xu, S., Chen, X., & Liu, D. (2009). Classifying software visualization tools using the Bloom's taxonomy of cognitive domain. In *Canadian Conference on Electrical* (pp. 13–18). <https://doi.org/10.1109/CCECE.2009.5090082>
- Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *CHI '23 Proceedings*. ACM. <https://doi.org/10.1145/3544548.3581388>

20 Publication 11. From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI

Memmert, L., Soroko, D., & Bittner, E. (2025). From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI. *Business & Information Systems Engineering*(67), 615–635. <https://doi.org/10.1007/s12599-025-00960-4>

Abstract

Generative Artificial Intelligence (GenAI) is adopted by knowledge workers to boost productivity, yet its specific characteristics such as probabilistic outputs and human-level content generation may change how professionals think about their effort. Prior literature has warned about unintended side effects of AI, but experiments on effort reduction when working with AI – which could threaten performance – reported mixed results. GenAI's rapid adoption combined with its specific characteristics make it critical and timely to clarify how GenAI influences knowledge workers' effort in professional settings. The qualitative study draws on 21 interviews with knowledge workers who frequently use GenAI for work. A directed content analysis, guided by expectancy theory and social loafing frameworks, revealed that most interviewees do not simply reduce effort, but rather strategically reallocate or even increase effort. They continuously learn to steer GenAI, viewing themselves as process administrators. The traditional group-based mechanisms of reduced effort or diffused responsibility do not seem to be directly transferable to human–GenAI dyads in professional settings. By revealing that GenAI reshapes the factors that influence effort rather than simply eroding motivation, providing a multifaceted view of effort investment beyond mere reduction, and highlighting the interplay between human relationships and GenAI-facilitated work, this research advances the discourse on human-(Gen)AI dynamics and the unintended consequences of (Gen)AI. Recognizing these shifts when setting policies and expectations enables organizations to benefit from GenAI's potential while mitigating potential risks to performance.

Keywords: Generative AI, GenAI, Human-AI collaboration, Social loafing, Smart loafing, Knowledge work

20.1 Introduction

Text-based generative artificial intelligence systems (GenAI) such as ChatGPT are a subset of artificial intelligence (AI) systems capable of producing content in the form of fluent text. While GenAI is being rapidly adopted (Hu, 2023) and can enable significant performance improvements for knowledge workers (Dell'Acqua et al., 2023), adopting AI may come with negative side-effects (Retkowsky et al., 2024). Studies suggest that working with AI might cause humans to reduce their efforts (Liu et al., 2023; Stieglitz et al., 2022), potentially “resulting in less-than-optimal team performance” (Liu et al., 2023, p. 2), if the AI cannot fully compensate for the reduction in human effort.

According to expectancy theory (Vroom, 1964), individuals adjust the amount of effort they put into a task based on their belief that this effort will yield desirable outcomes.

When **working alone**, individuals calibrate their effort to the perceived likelihood that additional effort will improve performance. Working with GenAI as compared to working alone fundamentally alters the task setting. GenAI can rapidly produce text and even creative content at a comparable level to many humans (Haase & Hanel, 2023), conquering aspects of tasks previously thought to be in the realms of humans (Dellermann et al., 2019). According to expectancy theory (Vroom, 1964), this change in the task setting can tempt individuals to consciously or unconsciously question the marginal benefit of additional personal effort, and potentially lead to a withdrawal of effort, if workers believe their input will not further improve an already proficient GenAI output (i.e., low expectancy).

When tasks are performed **collectively in human groups** that produce a single group result, individuals may also reduce their effort, e.g., because responsibility diffuses, or because their personal contribution seems less critical (Karau & Williams, 1993; Pinsonneault et al., 1999). The reduction of individual effort when working collectively is a phenomenon known as social loafing in human groups, which can decrease group performance (Karau & Williams, 1993; Latané et al., 1979). Social loafing has further been formalized as part of the collective effort model (Karau & Williams, 1993). Prior literature raised the question if humans show similar effort reduction behavior – akin to social loafing in human groups – when working with AI (e.g., Cymek et al., 2023; Hamza et al., 2024; Liu et al., 2023; Onnasch & Panayotidis, 2020; Rix & Hess, 2022; Siemon et al., 2015; Stieglitz et al., 2022). Observations that humans can perceive technical systems as social entities (Nass & Moon, 2000), and that – when working with an AI – they might shift responsibility to the AI (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022), suggest potential commonalities of human–human and human–AI dyads. Experiments in prior literature investigating non-GenAI systems, however, show mixed results, e.g., some concluding that social-loafing “occurs” when working with AI (Stieglitz et al., 2022, p. 759) while others find no effect (Siemon et al., 2015).

Based on the above, we argue that GenAI’s distinct capabilities and interaction modalities may lead humans to reconsider their effort when working with GenAI, which could lead to “less-than-optimal” performance or introduce risks. In the light of the rapid adoption of GenAI in professional contexts and the mixed results of prior laboratory studies it is, therefore, important to understand how knowledge workers’ effort changes when using GenAI in professional work contexts. We, therefore, pose the following research question: *How do knowledge workers manage their effort when working collectively with GenAI in a professional context?*

To answer this research question, we conduct a qualitative study applying a directed content analysis approach (Hsieh & Shannon, 2005). Drawing on previous research into changes in effort and loafing-like behavior in humans when working with AI, we have developed a conceptual model incorporating constructs such as expectancy, perceived responsibility and work performance. We conduct 21 interviews with knowledge workers who use GenAI in professional settings. The constructs in our literature-based conceptual

model serve as theoretical framework for analyzing the interviewees' reported effort investment rationales. We summarize the findings within these categories into subcategories to enhance clarity on how GenAI affects effort investment.

Our results illustrate several ways in which GenAI changes humans and their behavior, revealing that professionals do not simply reduce their effort, but strategically invest and reallocate effort to increase efficiency both on task- and job-level. More specifically, we explore how known antecedents for (reducing) effort are (re)shaped when working with GenAI, i.e., how GenAI may change human's perception of how (additional or reduced) effort translates into outcomes. For example, although for some tasks knowledge workers reduce their effort when perceiving a GenAI-based version to be sufficient (signaling reduced *expectancy*), they continuously learn to effectively steer GenAI (signaling high *expectancy*). Instead of an across-the-board erosion of motivation due to AI (Liu et al., 2023), this suggests that motivational antecedents are reshaped. We find that most workers feel fully responsible for the results created with GenAI, contrary to findings of earlier studies where participants shifted responsibility to the AI (e.g., Henkenjohann & Trenz, 2024; Stieglitz et al., 2022). Moreover, in line with earlier studies, we also find reduced effort for specific (sub-)tasks (Stieglitz et al., 2022), but our analysis suggests that workers tend to reinvest the saved effort into other (sub-)tasks. Lastly, we show that working with GenAI may affect relationships with others and vice versa, e.g., when creating work results with GenAI changes how others view and reward such results. Our analysis suggests that while established (collective) effort and social loafing perspectives offer a helpful lens for understanding how decisions concerning effort are made, describing the knowledge workers' behavior as 'loafing' might be a mischaracterization.

With our work, we contribute to the discourses on human-GenAI dynamics (Makarius et al., 2020; Memmert & Bittner, 2022), on the transferability of knowledge from human groups to human-(Gen)AI settings (Rix & Hess, 2022; Stieglitz et al., 2022), and on unintended consequences of introducing (Gen)AI (Memmert et al., 2023; Retkowsky et al., 2024). Specifically, we contribute an overview and testable propositions on how GenAI may reshape knowledge workers' motivation and effort in professional settings and discuss the potential influence of contextual factors. Building on prior conceptual work (Stieglitz et al., 2022), we suggest a multi-dimensional, multi-level *effort investment perspective* to reflect the inherent specifics of working with GenAI in professional settings, such as active prompting, probabilistic output requiring verification, and re-allocating effort to various (sub-)tasks. Our work has implications for knowledge workers, their managers and organizations, highlighting the need to acknowledge this new way of working when interpreting professionals' efforts and work results.

20.2 Background

20.2.1 Foundations of Effort Investment Research

Effort may be defined as the “conscious exertion of power or hard work” (Merriam-Webster 2024), but is best conceptualized as a multidimensional construct comprising the three facets of effort *intensity*, *persistence*, and *direction* (van Iddekinge et al., 2023); accordingly, van Iddekinge et al. (2023) define effort as “how hard workers try to perform their jobs, which includes where they devote their effort (direction), the amount of their effort (intensity), and how long they persevere in their effort (persistence)”. Notably, while intensity and persistence describe how work is performed, direction reflects what work is performed – a distinction informing our comprehensive conceptualization of effort for this study.

Expectancy theory (Vroom, 1964) proposes that individuals adjust their effort based on their subjective *expectation* or *belief* that additional input will yield valued outcomes. This theory posits that motivation – and, by extension, effort investment – is determined by three components: (1) expectancy, the belief that additional effort will enhance performance (Effort → Performance); (2) instrumentality, the belief that improved performance will lead to better outcomes (Performance → Outcomes); and (3) valence, the desirability or value of those outcomes. Consequently, effort decisions hinge on task context and the individual perception of whether additional effort exertion is worthwhile.

The Collective Effort Model (Karau & Williams, 1993) extends expectancy theory to group settings. Working with others instead of alone alters the task setting. In groups, performance is determined not only by the focal individual's effort but also by the contributions of other group members, and shared rewards may not align with personal goals. These contingencies in collective group work weaken the perceived link between one's effort and outcomes, altering motivations to invest effort. How such contingencies map onto GenAI work is examined later.

Working collectively means “individuals work in the real or imagined presence of others with whom they combine their inputs to form a single group product” (Karau & Williams, 1993, p. 681). If humans reduce their effort when working collectively rather than individually, this phenomenon is referred to as social loafing (Karau & Williams, 1993). Reasons for social loafing include the perception that one's effort is less crucial for group success when working with more proficient group members (dispensability of effort) or a reduction in individual responsibility (diffused responsibility) (Latané et al., 1979; Pinsonneault et al., 1999). This effort reduction may be conscious, following a deliberate appraisal of the task setting, or unconsciously, driven by “preexisting effort scripts” (Karau & Williams, 1993, p. 685). Importantly, individuals adjust their effort not based on the actual effort-outcome-relationship but according to their belief or perception thereof (Karau & Williams, 1993). Interestingly, these changes in effort are not only observable in actual effort exerted,

but already in the intention to exert effort (Hüffmeier et al., 2013). Factors such as group size modulate loafing risk, which can emerge even in dyads (Cymek, 2018; Cymek & Manzey, 2022).

Prior research applied expectancy theory (Vroom, 1964) as a framework for understanding motivation across diverse contexts, and social loafing has been documented in both physical and mental work settings (Karau & Williams, 1993). Operationalizing (or measuring) effort, however, is not straightforward. Previous studies have employed both direct measures (e.g., the number of calls made by a sales agent) and indirect approaches, such as self-ratings or ratings by peers and superiors (van Iddekinge et al., 2023). van Iddekinge et al. (2023) pointed out that while some studies have equated effort and performance, the two constructs are correlated, but distinct.

20.2.2 Effort Investment in Sociotechnical Systems

Text-based GenAI systems – pre-trained on large text corpora and engaging users conversationally (e.g., ChatGPT) – are rapidly diffusing across organizations, enhancing performance in knowledge-intensive tasks (Dell’Acqua et al., 2023). As GenAI becomes increasingly capable – producing creative ideas that rival human performance (Dellermann et al., 2019; Haase & Hanel, 2023) – humans may perceive their own effort as less essential (i.e., perceived *dispensability of effort*). Although working with GenAI differs from working with another human, prior literature indicates that certain dynamics observed when working collectively with AI toward an *integrated work product*, can resemble those found in human groups (Memmert & Bittner, 2024; Memmert et al., 2024; Rix & Hess, 2022; Stieglitz et al., 2022). Introducing (a probabilistic) GenAI weakens the effort-outcome relationship, because the work result becomes contingent upon the GenAI (to produce relevant output), analogous to the contingencies outlined in the collective effort model (Karau & Williams, 1993). Humans may view technical systems as social actors (Nass & Moon, 2000) or teammates (Liu et al., 2023; Rix, 2022; Rix & Hess, 2022; Seeber et al., 2020), specifically ChatGPT (Hamza et al., 2024), and studies suggest that humans shift responsibility to AI systems (i.e., *diffused responsibility*; Henkenjohann & Trenz, 2024; Stieglitz et al., 2022). Moreover, with human and GenAI being integrated into single work product, human and GenAI contributions may be difficult to discern (Retkowsky et al., 2024). In human group social loafing literature (Karau & Williams, 1993; Latané et al., 1979), the above factors are recognized as key antecedents of reduced individual effort (i.e. social loafing); and Rix and Hess (2022, p. 13) suggest that “due the intertwined nature of [human–machine teaming], it is harder to disentangle machine and human conduct, fostering the emergence of social loafing.”

Researchers suggested examining whether social loafing-like behavior occurs in dyads of humans and technical systems, involving “AI-like” systems (Siemon et al., 2015), virtual agents (Stieglitz et al., 2022), robots (Cymek & Manzey, 2022), or other AI tools (Liu et al., 2023), thereby extending the social loafing concept beyond human groups. Findings

are mixed: some report no effort reduction (Siemon et al., 2015), while others documented self-reported effort reduction, suggesting that as systems are increasingly seen as “legitimate and equal team members” (Stieglitz et al., 2022, p. 757), humans may reduce their own effort, leading Stieglitz et al. (2022, p. 759) to conclude that social loafing “occurs” when working with virtual agents. As for perceived responsibility, some studies find no difference (Cymek & Manzey, 2022), whereas others found a shift to the AI (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022). The mixed results may reflect measurement challenges for effort (van Iddekinge et al., 2023) and heterogeneity of technical systems, but render whether social loafing-like behavior occurs when working with GenAI uncertain. Moreover, whether social loafing is the right term is debated, particularly when AI compensates for reduced effort, leading Stieglitz et al. (2022, p. 758) to reframe it as “smart loafing,” where humans strategically reduce effort to conserve cognitive resources and enhance work efficiency.

Even if not fully subscribing to the view that GenAI induces social loafing per se, expectancy theory (Vroom, 1964) predicts that any shift in task structure – such as the integration of GenAI – will recalibrate individuals’ subjective evaluations of whether additional effort is worthwhile. Because GenAI can take over facets of knowledge work, users may question whether their contributions meaningfully improve outcomes and thus may adjust – or even reduce – their effort when working with GenAI. Importantly, reduced human effort does not necessarily translate into lower overall performance: GenAI’s efficiencies may offset or even surpass baseline outputs. However, reduced human effort can leave AI potential untapped or yield “less-than-optimal team performance” (Liu et al., 2023, p. 2). Accordingly, we investigate how GenAI shapes knowledge workers’ effort investment and revisit whether social loafing is an appropriate lens for GenAI work.

20.3 Derivation of the Conceptual Model

Decades of research on both expectancy theory (Vroom, 1964) and social loafing (Latané et al., 1979) across domains and across both mental and physical work demonstrates that individuals adjust their effort based on subjective perceptions (Karau & Williams, 1993; van Iddekinge et al., 2023). As GenAI becomes increasingly capable and may be perceived as partners when performing knowledge work collectively (Hamza et al., 2024; Rix & Hess, 2022), they may change workers’ perceptions of the impact of their own effort. To analyze how GenAI might shape knowledge workers’ effort investment decisions – directly and through its antecedents – we propose a conceptual model as a theoretical framework (see Figure 1), which we have derived from literature as described below.

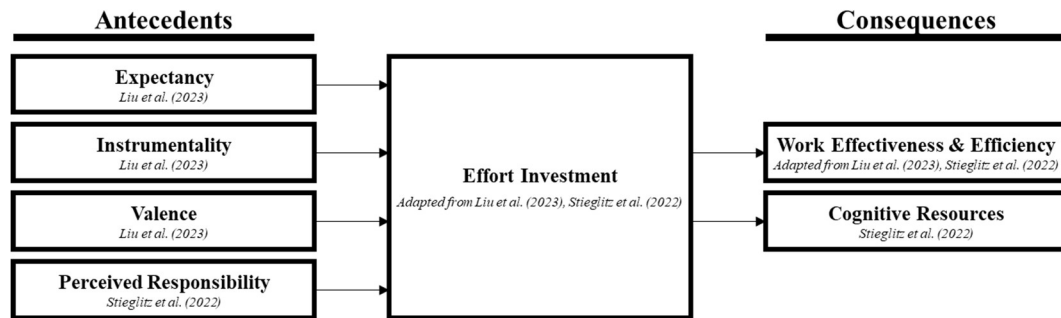


Figure 1. Conceptual model derived from Liu et al. (2023) and Stieglitz et al. (2022)

Literature on expectancy theory and social loafing is vast (van Iddekinge et al., 2023). To ensure relevance with respect to the emerging discourse on social loafing-like behavior and effort management when humans work with technical systems, we derive our model from the recent works of Stieglitz et al. (2022) and Liu et al. (2023), as these studies, too, discuss changes in effort investment with conversational or AI systems in one-on-one settings (i.e., one human working with one system) when performing knowledge work. From these papers, we extract the discussed antecedents as well as the consequences of effort reduction when working with AI. We include factors that are relevant to effort investment and consolidate and adjust them for conceptual clarity. We exclude fixed personal traits (e.g., personality factors) that are less likely to immediately shift when working with GenAI, as we sought to understand the impact of GenAI on knowledge workers' effort investment – though we acknowledge that personal differences likely affect those decisions. Each construct is briefly discussed below and defined as a foundation for our analysis of the interviews with knowledge workers. While for the selection of the antecedents and consequences (e.g., defining the scope of the analysis), we used the works of Stieglitz et al. (2022) and Liu et al. (2023) as a foundation to ensure relevance to the emergent discourse on collective human-AI knowledge work (see rationale described above); for the definition and explanations of those constructs, we used literature on expectancy theory and social loafing more broadly, to ensure consistency with prior literature.

For antecedents – following Liu et al. (2023) – we include *expectancy*, *instrumentality*, and *valence*, strongly rooted in expectancy theory and the collective effort model (Karau & Williams, 1993). Expectancy refers to the belief that additional effort will enhance performance; instrumentality to the belief that improved performance will lead to better outcomes; and valence to the desirability or value the human places on these outcomes (Karau & Williams, 1993). Additionally, we include *perceived responsibility* for the work outcomes as an antecedent following Stieglitz et al. (2022), who discussed the attribution of responsibility between the human and the technical system. We did not explicitly include expertise (as mentioned by Stieglitz et al., 2022), as this is subsumed under expectancy, i.e., the expectation of the degree of personal capability compared to the AI capability (also see discussion in Liu et al., 2023).

The central construct in our conceptual model is effort investment. While Stieglitz et al. (2022, p. 761) operationalized human effort via “social loafing tendencies in virtual collaboration with the virtual assistants”, i.e., with a effort-reduction-perspective, we adopted a more neutral perspective on effort, more akin to the conceptualization of Liu et al. (2023, p. 5), who referred to it as “human effort in human-AI team.”

For consequences of changed effort investment, we included work effectiveness and efficiency as well as cognitive resources, as these are the central concepts in the “smart loafing” construct developed by Stieglitz et al. (2022, p. 758), which they define as “the reduction of effort in human-[virtual assistants] collaboration to maintain *cognitive resources* and enhance *efficiency in work*” (emphasis added). *Work effectiveness* and *efficiency* refers to how the work process and outcomes are affected by the changed effort. Following the smart loafing perspective (Stieglitz et al., 2022), for *cognitive resources* we refer to the personal mental capacity being conserved or depleted while working on the task, e.g., due to increased or reduced cognitive load (Sweller, 1988). Though cognitive resources could be included into efficiency, we decided to differentiate the two constructs due to the difference in the referend, i.e., task versus individual.

We do not challenge the underlying theories, which have been tested across many domains and contexts. With our model, we integrate a sub-set of antecedents and consequences contained in the model of Liu et al. (2023) and the smart loafing definition of Stieglitz et al. (2022), to guide the investigation of how GenAI shapes motivation and effort investment decisions of knowledge workers in organizations (Table 1).

Table 1. Literature-based category system used as deductive frame with categories and explanations

Category	Explanation	References
Expectancy	Refers to thoughts about the believe of the ability to improve task performance, as lower expectancy can lead to effort reduction, whereas the feeling of being capable to make a unique contribution might make effort reduction less likely to occur	Karau and Williams (1993); Liu et al. (2023)
Instrumentality	Refers to thoughts on the necessity to contribute to achieve better outcomes through improved performance, as lower perceived instrumentality (i.e., a higher perceived dispensability of one's effort) can lead to a reduction in effort	Karau and Williams (1993); Pinsonneault et al. (1999); Liu et al. (2023); Memmert and Tavanapour (2023)
Valence	Refers to thoughts on how valuable or desirable the knowledge worker perceives the outcome of the tasks to be, with high valence being associated with a reduced likelihood of reducing effort	Liu et al. (2023); Williams and Karau (1991); Karau and Williams (1993)
Perceived responsibility	Refers to thoughts on the responsibility for work results, as (perceived) responsibility diffusion might lead to a reduction in effort (responsibility attribution towards technical systems occurred in prior literature)	Pinsonneault et al. (1999); Stieglitz et al. (2022); Henkenjohann and Trenz (2024); Memmert (2024)
Effort investment	Refers to thoughts about how to calibrate the own effort invested in the task, excluding general strategies on how to work effectively with GenAI, like prompt engineering	Stieglitz et al. (2022); Memmert (2024)
Work Effectiveness and Efficiency	Refers to thoughts and the reasoning on leveraging GenAI for the tasks, as enhancing work efficiency was highlighted as a key reasoning for leveraging technical systems	Stieglitz et al. (2022); Liu et al. (2023)
Cognitive Resources	Refers to thoughts about the management of and effect on mental resources when working with GenAI	Sweller (1988); Stieglitz et al. (2022); Hart and Staveland (1988)

20.4 Method

To investigate how professionals determine their effort investment when using GenAI in their daily work, we used a qualitative approach, interviewing 21 knowledge workers. The interviews took place between March 2024 and March 2025. The interviewees were recruited via a professional society newsletter as well as the authors' personal network. We selected only employed professionals to understand their effort investment decisions in a work context embedded in organizational structures, where the importance of the task outcome is representative for work settings and matters to other stakeholders. When selecting participants, we ensured that all interviewees met the established criteria for knowledge work, including white-collar roles (Mandt, 1978, as cited in Timonen & Paloheimo, 2008), low standardization (Pyöriä, 2005), engagement with abstract knowledge and symbols (Pyöriä, 2005), reliance on specialized skills and theoretical knowledge (Timonen & Paloheimo, 2008), non-routine tasks (Järvenpää and Eloranta 2001, as cited in Timonen & Paloheimo, 2008), problem-solving abilities (Tsoukas and Vladimirou 2001, as cited in Timonen & Paloheimo, 2008), and the need for extensive formal education and continuous on-the-job learning (Pyöriä, 2005). Our sample also included knowledge workers who had experience using GenAI, such as ChatGPT, in their daily work. We used a broader sample

to get a more general understanding of how knowledge workers think about the effort investment and not be biased due to a specific type of profession. We have listed the industries, job titles and overall professional experience in Table 2.

Table 2. Interviewee details

ID	Job	Professional Area	Professional experience range [years]	Interview length [mins.]
I1	Manager	Energy Infrastructure	5..8	38
I2			5..8	33
I3			5..8	41
I4		Consulting	9..12	46
I5			1..4	39
I6		Banking	9..12	31
I7		Industry Park	9.12	35
I8	Leader and Manager	Parliamentary Group	5..8	32
I9	Consultant	Consulting	1..4	45
I10			5..8	39
I11			>12	42
I12	Associate or senior researcher	Research	1..4	42
I13			5..8	44
I14			5..8	65
I15			5..8	58
I16			>12	48
I17			1..4	33
I18			>12	52.
I19			5..8	63
I20		Automotive	5..8	50
I21			5..8	33

The interviews were semi-structured (Meuser & Nagel, 2009), following an interview guideline developed along our research goal and conceptual model (Fig. 1), but leaving room for personal experiences of the professionals. The guideline underwent a slight revision after the first 16 interviews to incorporate more detailed clarifying questions (see Appendix A, available online via <http://link.springer.com>). After providing a general introduction to the study and obtaining consent, we began the first part of the interview by asking about interviewees GenAI usage in professional work, and – while not at the core of our study – about the tasks they use GenAI for. This allowed us to build rapport with the interviewees, preparing them for sharing more detailed reflections in the interview, but also allowed us to contextualize the results. In the second part, we asked participants to reflect on their effort investment when using GenAI. Depending on the answers, we asked follow-up questions, e.g., regarding the interviewees' first steps when working jointly with GenAI on a task, routines they had developed while using GenAI, how they determined when to stop working together with GenAI on the task, and how they integrated the outputs into their final work results. In the third part, we asked participants to reflect on how the work

results produced with GenAI felt to them personally, and how they felt working with GenAI affected their mental workload and stress. Interviews were closed with an open-ended question on whether participants wanted to share any additional information. We made sure to ask open-ended questions as well as indirect questions to prevent biasing the respondents (Döring & Bortz, 2016). To make the interviews concrete, we asked participants to share specific examples (if they were allowed to disclose the information) or even share their screen to walk us through their conversations with GenAI. We used this approach to invite participants to reflect on their behavior, because – as was pointed out in the background section – humans may sometimes adjust their effort unconsciously within certain situations and might become only aware when actively reflecting on it.

All interviews were conducted online, were recorded with the participants' permission, and were later transcribed verbatim. We followed a qualitative content analysis approach (Hsieh & Shannon, 2005; Mayring, 2014), more specifically, a directed content analysis (Hsieh & Shannon, 2005). Hsieh and Shannon (2005, p. 1281, *emphasize added*) explain that the “goal of a directed approach to content analysis is to [...] *extend conceptually a theoretical framework* [...]”, which is helpful if “*existing theory or prior research exists about a phenomenon* that is incomplete or would benefit from further description”. This approach fits well with our research goal, because established theories exist (i.e., expectancy theory, collective effort model) explaining effort changes in human groups, which underwent initial extension to work with AI-based systems and were integrated into a *theoretical framework* (see Fig. 1). We now seek to explore and *describe* how human effort changes when working with GenAI. Following Hsieh and Shannon (2005, p. 1281) we begin by using “existing theory or prior research” to identify “key concepts [...] as initial coding categories” to guide our analysis and add “operational definitions for each category”; see Fig. 1 with detailed explanation in Sects. 2 and 3, and Table 1 with the categories and definitions. We then proceed to code the interviews (appendix B contains the full coding schema). Individual statements with complete arguments were used as primary coding units. As suggested by Hsieh and Shannon (2005), we grouped our findings into sub-categories to enhance clarity. Two interviewers first coded a small sample of all interviews (two interviews each) to shape the initial coding scheme. This was followed by a round of discussions to resolve any disagreements and inconsistencies between the coders. Then the same sample of interviews was re-coded individually once more to align data analysis to the revised coding scheme. Finally, the rest of the interviews were analyzed according to the established coding guidelines. To make the results more tangible, we selected direct quotes from the interviews, which were translated for interviews not conducted in English, to be included in the final manuscript.

After coding the initial set of 16 interviews, we reached data saturation, defined as the point at which no new (sub-)categories or insights were emerging from the collected data. Saturation was assessed by both coders through ongoing discussions throughout the data analysis by reviewing transcripts and developed (sub-)categories to ensure that they adequately

captured the data. The subsequent five interviews did not yield any new information and confirmed earlier findings. This is consistent with prior research suggesting that data saturation in qualitative studies can be reached as soon as twelve interviews (Guest et al., 2006).

20.5 Results

To contextualize the interviewees' reflection on working with GenAI, we first asked them about which GenAI they used, and how frequently and for what kind of tasks they used GenAI. We find the interviewees mostly use OpenAI's ChatGPT and some use Microsoft Copilot. Many interviewees use GenAI multiple times daily and all use GenAI for their professional knowledge work. This was expected given our research goal and sampling criteria. Our interviewees cover a variety of tasks, from low- to high-stakes and from operational to strategic tasks. Examples include summarizing notes (e.g., I10), formulating, grammar checking, or editing emails and texts (e.g., I12-15), translating documents (e.g., I9-I10, I14), sanity-checking concepts or presentations (e.g., I8, I21), developing user stories (e.g., I10-I11), developing training material and exercises for students or fellow employees (e.g., I6, I7, I10, I14, I16) or developing concepts (e.g., I9, I11, I18). Most of the tasks met our knowledge work criteria mentioned in chapter 4. The reported "conversation lengths" with the GenAI (i.e., number of messages sent by the user) differed from one-off interactions up to week-long conversations.

We have structured the results section according to the conceptual model presented earlier, with effort investment, and potential antecedents and consequences. We have summarized the interviewees' considerations reported in Figure 2, which may increase, retain, decrease certain model elements; we have indicated the *tendencies* according to the interview data, or flagged them as ambiguous. However, these need to be treated as preliminary and with caution, requiring (quantitative) validation.

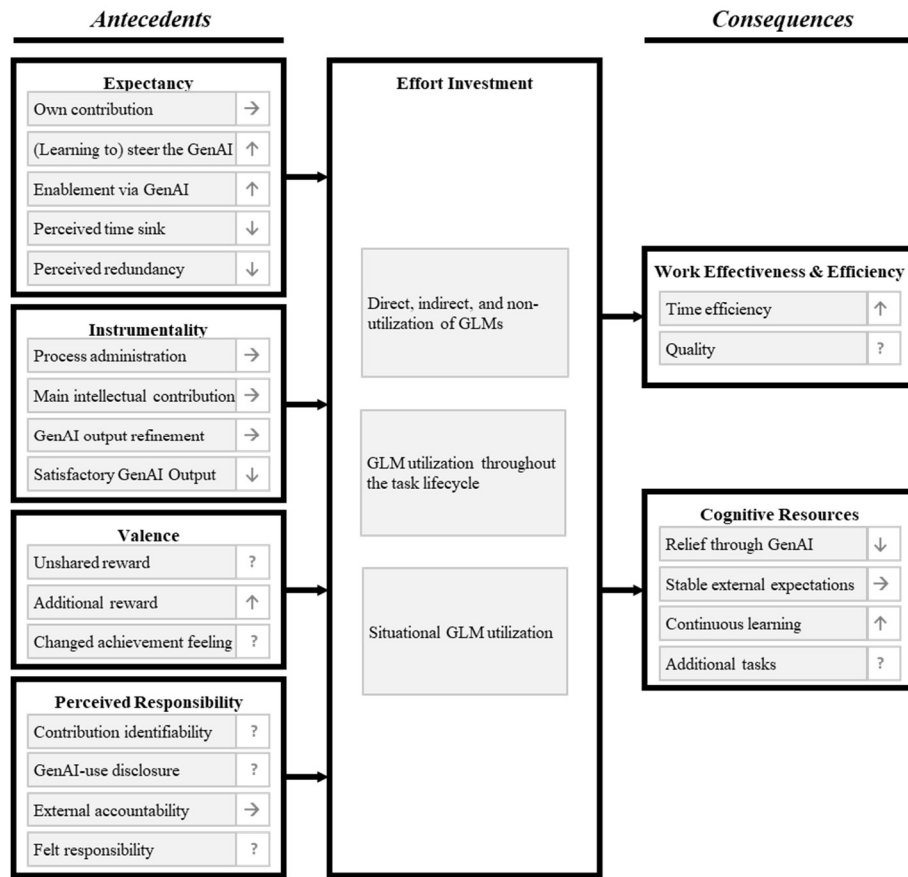


Figure 2. Changes to knowledge workers' effort investment when working collectively with GenAI.

Note: We assessed the reported influence of working with GenAI on antecedents and consequences of effort investment; arrows indicate the reported direction – increase (↑), retained (→), decreased (↓), or ambiguous (?); direction should be interpreted with caution and require further (quantitative) validation

20.5.1 Expectancy

Expectancy refers to the human's belief that (additional) effort will improve performance (Vroom, 1964). Our results show that – when working with GenAI – this can involve not only the human's own direct contributions but also their perceived ability to steer the GenAI towards creating better outputs, and even enablement to work on tasks that the human could not have performed alone.

Own contribution (→) Generally, participants' responses reflect a high sense of expectancy with most participants reporting that they believe they can make own, direct contributions that improve performance. Specifically, many participants noted that they provide overall *strategic or structural knowledge* for the task (e.g., I11, I18-I20). Additionally, they report that they provide the *situation-specific, expert knowledge* (e.g., I17, I18-I20) required for most tasks, that cannot be generated by GenAI, with one participant explaining “the results [produced by the GenAI] are semi usable if you don't put your own know-how into it [...]. What we do requires a certain creative approach, which the systems currently do not yet provide” (I4).

(Learning to) steer the GenAI (↑) Beyond making direct contributions, participants also reported investing up-front and continuously into learning to effectively steer GenAI. This includes, e.g., reading blog articles and prompting guides as well as reviewing co-workers' conversations with the GenAI (e.g., I11) and attending workshops (e.g., I16). One interviewee (I5) explained “especially at the beginning, as I said, I looked a lot at blog articles and things like that on how you can, may, should optimize your workflows with [GenAI]; what kind of prompt is suitable or whatever. But now I think I know that I at least have a set of things that work quite well for me.”

This covers operational, conceptual, and procedural aspects. On an operational level, interviewees try to learn how to effectively prompt, e.g., deliberately providing context information, certain perspectives or keywords, specifying guardrails or frameworks to follow (e.g., I9, I19). They calibrate the prompt complexity for the task complexity (e.g., I3, I9). Beyond prompting design techniques, some learn and re-use prompts for recurring tasks (e.g., I10, I13-I16) or have even developed prompts library (e.g., I3, I15). Conceptually, interviewees test and try to understand the boundaries of GenAI (e.g., I9, I16, I20), and adjust their usage behavior accordingly, e.g., by editing prior messages in conversations to remove unneeded information (e.g., I9), or by fact-checking GenAI outputs. One participant (I20) stressed the importance of having an idea for how GenAI functions to be able to delegate tasks appropriately. On a procedural level, the focus lies on understanding how to integrate GenAI into their own workflow (e.g., I5, I15, I19).

Learning how to effectively utilize GenAI is not perceived as a one-time effort, but as a continuous learning activity (e.g., I7-I8, I11, I15-I16), requiring effort to improve prompting efficiency, advancing the prompt library, and monitoring the evolving GenAI tool landscape with different GenAI systems (e.g., I9, I10, I14). This sustained effort investment illustrates participants' belief that in-situ steering or learning to better steer the GenAI leads to greater control over the model's output and, in turn, meaningfully improves performance. This belief of exerting effort in the interaction or even investing effort up-front to increase performance, signals a high level of expectancy (Vroom, 1964).

Enablement via GenAI (↑) Participants report to be enabled to perform new tasks they otherwise could not have performed. Thus, while participants might have felt low expectancy with respect to certain tasks pre-GenAI, due to GenAI, their skills are elevated making their effort productive towards achieving new outcomes, e.g., working on a task with aspects outside their expertise (e.g., I6, I17-I18). An interviewee recently started learning to code, and with the help of ChatGPT they feel capable of tackling most of the arising coding problems (I9). Others explained, the GenAI allows them to focus on the structuring and creative part of the task (e.g., I5), or that they felt “powerful” when having access to ChatGPT: “At the fingertips, I have so much power now that I can use!” (I12).

Perceived time sink (↓) Signs for reduced expectancy were rare. While most participants believe they can (learn to) control GenAI, some participants pointed out that using a GenAI

is sometimes perceived as a time sink, when they cannot get the GenAI to produce the output they need for certain tasks (e.g., I16), with one participant explaining “I spend so much time now essentially arguing with GPT about a solution that I'm not happy with. And this was really not fruitful in any way and kind of a waste of my time” (I14). This shows a parallel to the contingency in human groups discussed by Karau and Williams (1993): when working with others, the relationship between own effort (input) and the work result (outcome) is less direct (as compared to when working individually), because the outcome is dependent on additional factors such as other group members' contributions. Similarly, when working with a GenAI, the outcome becomes contingent upon the GenAI to produce relevant output. When professionals feel that interacting with GenAI burns time without improving its output, their expectancy – the belief that extra effort will pay off – is reduced. Overall, GenAI is not viewed a waste of time, but as a means to enhance productivity. Interestingly, several participants perceived working with GenAI as positive even when not receiving the expected results, using it as a learning opportunity.

Perceived redundancy (↓) However, some participants pointed out that for some parts of their task like phrasing or expressing thoughts in another language, they perceive the GenAI as more proficient, rendering their contribution – for this aspect of the task – redundant or irrelevant (e.g., I17), signaling reduced expectancy.

20.5.2 Instrumentality

While expectancy focusses on whether additional effort will improve performance, instrumentality refers to the belief that improved performance leads to better outcomes (Karau & Williams, 1993; Vroom, 1964). When group members feel that their effort is less instrumental (i.e., feel their effort to be dispensable) for group success, they might reduce their own effort. In the prior subsection, it was reported that participants generally felt that despite GenAI becoming more capable, participants' effort would lead to enhanced performance (expectancy). Beyond expectancy, most participants felt that their contribution not only enhanced performance but was necessary to achieve the work results that stakeholders expect. Hence, their effort is *instrumental* in achieving better outcomes, which indicates sustained instrumentality despite the GenAI. Participants explained this with respect to the way GenAI is integrated into their job more fundamentally (see *process administering*), the nature of participants' contributions (see *main intellectual contribution*), and the GenAI's output quality. Only in rare cases was the GenAI's output quality sufficient on its own, indicating a potential reduction in instrumentality.

Process administration (→) We asked participants to reflect on how they integrate GenAI into their work. Participants reported that they understand the task and expectations, enter the required information into the GenAI, iteratively develop, evaluate and refine GenAI outputs, and shape them according to stakeholder expectations. Because participants still orchestrate the workflow and integrate the knowledge into the organization – leaving the

core of their work untouched (I5) – their effort remains indispensable for achieving the desired outcome, signaling unchanged instrumentality.

Main intellectual contribution (→) On a more detailed level, participants point out that usually it is them who contribute the main intellectual, structural, or strategic knowledge, while GenAI serves more as an editor (e.g., I13, I17). Hence, while GenAI can support with many task aspects, without the participants there would not be a sufficient foundation for the GenAI to produce work results in the first place. Interviewees felt able to make unique contributions, e.g., in the form of their expert knowledge and experience (I10, I12-I14, I19-I21), whereas unique contributions by the GenAI were rare (e.g., I6, I11, I19). Supplying the strategic knowledge that the GenAI lacks keeps human effort instrumental to the final outcome.

GenAI output refinement (→) On a more operational level, participants state that they cannot use the output as-is, explaining that it is usually too general or contains false information, resulting in them having to fact-check, refine, and enrich the output with their own expert knowledge (e.g., I5, I7-I8, I19-I20). Interviewees assure quality with varying degrees of effort investment by them and others. In more general terms, interviewees stated to have a certain quality-related “self-image” of their work (e.g., I6, I9, I13-I14), both personally and for their company towards their clients, or as one interviewee (I5) put it “It's also somewhat of a principle of our work to only send out high-quality results”. When engaging with the GenAI to work on a task, interviewees reported to have developed expectations regarding the results upfront (e.g. I2-I3, I6, I18-I19). Given participants generally seem to be aware of (some of) the limitations of GenAI – e.g., naming hallucinations (e.g., I3) – they report to critically review and edit the GenAI's output based on their own (expert) knowledge and experience (e.g., I4, I7-I8, I12-I15, I19-I21), while also considering stakeholder expectations (e.g., I2), and to verify the outputs with original sources (e.g., I5, I18-I20); or, as one interviewee (I3) put it: “In my opinion, we will never be spared the step of saying that a human who knows what they're doing needs to look at [the GenAI output] again to say whether it fits or not.” From these explanations it becomes clear that fact-checking and enrichment is perceived to be necessary, meaning the task cannot be completed without human intervention, sustaining instrumentality.

While stressing having checked the GenAI's output, interviewees revealed instances in which they partially outsourced quality assurance. One interviewee (I6) shared that they had to create a document in a field outside their expertise. Given this was a one-time task and familiarizing oneself with the relevant literature would have meant a large effort investment, the interviewee instead leveraged a GenAI to produce the document and sent it to a different department, which holds the final accountability. A different approach that was mentioned was to clearly disclose to have used GenAI to manage expectations that the output might be taken with a grain of salt, or might not be helpful (I2, I10).

Satisfactory GenAI output (↓) While participants generally indicated a high sense of instrumentality and usually refine GenAI-produced output, several participants pointed out that for some tasks as part of their jobs in the organization the GenAI output or “80%-solutions” are sufficient (e.g., I2). This means that the humans’ effort would not be required anymore for those tasks because – even though humans could improve performance (expectancy) – such increased performance would not lead to better outcomes (because GenAI outputs already meet organisational standards), indicating a reduction in instrumentality for these aspects of their job.

20.5.3 Valence

Valence refers to how valuable the human finds the outcome when completing a task (Karau & Williams, 1993; Vroom, 1964). Compared to individual work, the connection between individual performance and individual outcomes are less direct in group settings due to additional contingencies (Karau & Williams, 1993). Rewards may be provided to the group but may not necessarily translate into valued outcomes for the individual, e.g., because the group gets praised, but the individual’s contribution is not recognized. We were curious if working with GenAI affected valence, i.e., if working with GenAI changed the value perception of the work results. However, results show that generally, participants feel similarly rewarded even when working with GenAI and might even receive more rewards. However, working with the GenAI can alter the feeling of achievement.

Unshared reward (?) With GenAI partially taking over, it might be conceivable that others, when evaluating the work results, reduce their praise or credit, because not everything was done by the human alone, analogous to human group settings, where praise may be divided among group members. However, participants reported that they generally do not face this problem (e.g., I18). Work results are usually perceived as theirs, hence, they receive the reward (i.e., unshared reward), indicating sustained valence. However, there were instances in which superiors joked about the contribution of the human (e.g., I10), given that a large part of the results was produced with GenAI. Thus, with even more widespread adoption, peers and managers might indeed start judging work results differently when they can be produced (to a large extent) with GenAI, potentially deferring praise partly to “the GenAI,” and reducing praise for the involved human.

Additional reward (↑) Additionally, participants pointed out that because they can perform tasks better; focus on more value-adding aspects of the tasks (e.g., I5), can perform more work in general being more productive (e.g., I5), and are even enabled by GenAI to do tasks they could have not done without it (e.g., I9), i.e., they can do more and potentially received additional rewards they might have not received otherwise.

Changed achievement feeling (?) Beyond the extrinsic, there is an intrinsic dimension to valence (Karau and Williams 1993). GenAI may affect how participants feel about (some of their) work, i.e., when developed jointly with GenAI, they perceive it as “less” or have a reduced feeling of accomplishment (e.g., I2, I9), which could reduce intrinsic motivation.

One participant (I9) explained it as follows “It feels a bit like someone else is solving the problem for you.” However, most participants explained that they weave the GenAI’s output into the work results (e.g., I18), reporting to feel it to be “their” work, e.g., stating “so after rewriting it a few times, I think it has become my text rather than a generated text” (I16). This point is closely connected to a feeling of responsibility (next section). One participant (I19) also reported creating a very particular routine when using ChatGPT for most of her tasks, which enabled her to “actually enhance the meaningfulness of the task” and made her feel much happier about the outcome.

20.5.4 Perceived Responsibility

An important factor influencing the level of effort is how responsible a person feels for the task and outcome (Karau & Williams, 1993; Pinsonneault et al., 1999). Literature on social loafing shows that making individual contributions in group work identifiable, can increase responsibility and reduce social loafing behavior (Karau & Williams, 1993). Whether or not recipients of the work results can identify the parts of the work results produced by the human versus by GenAI, depends on whether (1) they can tell the GenAI and human contribution apart, or, if they cannot, on whether (2) the person sharing the work results tells the recipients if and how they have used GenAI when producing the results. We inquired participants on both questions. Additionally, as prior literature showed that humans might shift responsibility to AI (Stieglitz et al., 2022), we also asked for the felt responsibility more directly.

Contribution identifiability (?) Responses on whether interviewees felt that the recipients of their work results can distinguish, which parts originated from them and which parts from GenAI were mixed, some said the contributions are indistinguishable (e.g., I3, I6-I7, I16, I20), e.g., because they worked iteratively and integrated the GenAI’s and their own thoughts closely in the process. Other participants were unsure (e.g., I5), or believed contributions to be distinguishable (e.g., I13-I15). Hence, the identifiability of contributions may depend on contextual aspects like usage behavior rendering the effect on perceived responsibility uncertain.

GenAI-use disclosure (?) When interviewees have the feeling their contributions and those by the GenAI are not distinguishable for the recipient, they have the choice of disclosing having worked with GenAI or not. Generally, interviewees seem not to disclose having worked with GenAI (e.g., I3, I5-I6, I12-I14, I20-I21), drawing a comparison with other tools like Google, which they also do not explicitly disclose to have used (e.g., I3, I8); explaining that “the part they generate is usually only a portion of what [they] send out” (e.g., I5), with the majority being written by them; that there is no organizational rule requiring such disclosure (e.g., I3); or that they also do not “explicitly claim it as [their] result” (I6). In contrast, reasons for disclosing having worked with the GenAI include promoting a new way of working and encouraging using GenAI for work (e.g., I10); working closely together with people (e.g., I6); or certain rules requiring disclosure (e.g., I12-I15,

I21). Additionally, a reason for disclosure might be not wanting to take false credit and managing expectations towards co-workers regarding one's ability, or as one interviewee put it: "Why should I claim something as mine, if it is not and something I don't do it well?" (e.g., I10). Lastly, as mentioned above, disclosure might also be used to help co-workers to better assess the quality of the results (e.g., I10). Choosing whether – and how – to reveal GenAI use gives workers control over identifiability, so responsibility effects may vary.

External accountability (→) After discussing known causes associated with reduced responsibility and social loafing behavior (Pinsonneault et al., 1999), we directly about participant's felt responsibility and ownership regarding the results: most if not all participants felt highly responsible. Reasons mentioned included being held accountable for the results and being judged by the outcomes of their work with expectations remaining the same (e.g., I6, I19) irrespective of how results are produced, or, as one interviewee put it: "nobody cares about the way the results are produced – it is perceived as coming from me" (I2).

These results surface an interesting difference between established knowledge on human groups and the human- GenAI dyads in our study when it comes to identifiability of contributions. For recipients of the work, it may not be easy to understand, which parts of the work results were actually produced by the human (and which by GenAI), and the creator of the work result has a choice of disclosing whether and how they used GenAI. Based on prior literature, non-identifiability and non-disclosure together could mean, that the actual contribution of the human is concealed, which might allow the human to reduce their effort, analogous to social loafing behavior. However, unlike in human groups, there are no other humans when working with GenAI. Because stakeholders treat the human as the sole producer, participants reported being held as responsible for the outcome as if they had not used the GenAI.

Felt responsibility (?) Regarding whether their work results produced alone felt different as opposed to when developed using GenAI, responses among interviewees were mixed. Some interviewees stressed they feel results produced with GenAI are their work (e.g., I3, I6, I12-I16, I20-I21), offering varying reasoning, e.g., that the main thoughts for the result still come from them (e.g., I11-I15, I18, I21), that they had written the prompts (e.g., I3, I11), or that they had invested time (e.g., I14). One of the most prominent points made was a sense of intellectual or creative ownership, with several participants stating that the main ideas and intellectual effort came from them (e.g., I17-I18), while GenAI was more of an editor or assistant tool (e.g., I12-I15, I21). Thus, being the main intellectual contributor and process orchestrator does not only lead to instrumentality (as reported earlier), but also to feeling responsible.

An additional perspective was provided by one interviewee (I6), who explained that being a manager they delegate many tasks anyways, so using GenAI would not change how they perceived the work results, stating "it is about the task" (I6). Other interviewees, however,

reported the results produced with GenAI felt different. Potential explanations offered included having to spend less time developing the results, and, thus, having felt less personal attachment (e.g., I1); that the results sounded more eloquent than the interviewee could have formulated it (e.g., I10), and that working with GenAI is a bit like taking a short-cut (e.g., I2) or “it feels a bit like someone else is solving the problem for oneself” (I9), which takes away from the feeling of achievement (e.g., I2, I9). This does not necessarily mean they deny responsibility, but they might feel less personal attachment. Hence, the felt responsibility may vary, e.g., depending on personal factors (such as how one views one own work) and on how GenAI is integrated into the work.

20.5.5 Effort Investment

We asked interviewees to reflect on their considerations towards managing their effort investment when working with GenAI, finding participants have developed certain patterns on when, how, and why they use GenAI, with implications for effort intensity, persistence, and direction.

Direct, indirect, and non-utilization of GenAI. Several interviewees stated to include large portions of GenAI’s output in the final work results, with estimates ranging from 50 to 85% of text copied (e.g., I3, I5-I6, I14), particularly for texts in which GenAI is used as a writing assistant. Against this backdrop, a participant (I17) reported to sometimes use a simpler writing style just to get something on paper, because they know GenAI will rephrase appropriately, explaining “before you don't write anything because you're thinking too much, I've gotten into the habit of just starting to write and then I can always read about it later or just tell the AI now, can you please go over my low effort here and work it up a bit linguistically”. Interviewees not only use the system directly for their work result, but also indirectly, e.g., to produce speaker notes for slides (e.g., I2), to get a better understanding of a topic (e.g., I5, I9, I15, I19-I20), or for internal knowledge organization more generally (e.g., I9). However, though helpful for many tasks, GenAI is not capable to support many other tasks in the interviewees’ jobs (e.g., I7-I8, I16) – interviewees deliberately choose for which tasks to engage GenAI, and when to perform the task alone.

GenAI utilization throughout the task lifecycle. Social loafing-like behavior might result in humans reducing their effort to let GenAI do their work. We, thus, asked the interviewees about how they used GenAI throughout the task lifecycle. Interviewees seem to employ GenAI throughout working on their tasks, from the very beginning, throughout performing the task, and in the end, with varying reasoning. In the beginning, interviewees use it as a replacement for doing their own research, to lower the entry barrier and quickly get an overview of a topic outside their expertise (e.g., I5, I7, I9-I10, I15, I18-I20), to quickly generate draft versions with just a few inputs (I5) instead of starting with a “blank paper” (e.g., I1, I14, I8), or to generally get inspiration on “where to go from here” (e.g., I6, I12-I13). When used throughout working on a task, interviewees use GenAI to get a different perspective (e.g., I10, I21) or as a “sparring partner” (e.g., I3, I6, I8-I9, I11, I18) to jointly

develop a solution incrementally. One interviewee explained: “When structuring, we use [GenAI] as double check in the sense of what does [the GenAI] suggest, or what can be optimized further” (I11). Many interviewees also use GenAI at the end of the task, such as to check grammar (e.g., I8, I12-I13), to help writing a text in a different language (e.g., I6, I9, I17), or to perform last sanity or completeness checks on a work product like a presentation (e.g., I11, I21).

Situational GenAI utilization. Maintaining cognitive resources is a core idea of smart loafing (Stieglitz et al. 2022). Interviewees (e.g., I5, I17) specifically reported using GenAI when feeling low in energy, lacking the mental capacity to tackle certain tasks, particularly in the hour before lunch and in the afternoon hours. They then used GenAI to create a first draft of the work product and return to it in times of higher energy. Starting to work on a new topic at the border or outside of one's own area of expertise can be time-consuming. Several interviewees have faced this issue and used GenAI to ease their start into the new topic when seeking to build up expertise (e.g., I5, I9, I15, I18-I19). One participant (I18) explained to that end: “Sometimes, when I'm under time pressure and I have a new topic where I don't yet have a complete overview or perhaps, I want to get into the technical details quickly and easily, I use ChatGPT as a bit of a sparring partner to get a better overview of the topic”. Another participant (I20) offered more nuance by saying that their use of GenAI depends on how sophisticated and complex the task is, the more sophisticated the task, the less likely they were to use GenAI.

If they need to perform a onetime task in which they are not an expert, they might also use GenAI so they do not have to build up the skills they would need to complete the task by themselves (I6, I10), e.g., to break down complicated, domain-specific KPIs or to apply legal regulations.

It becomes apparent that participants deliberately choose how, when, and why to engage GenAI, with consequences to effort intensity (e.g., trying less hard for tasks which can be covered by GenAI), effort persistence (e.g., overcoming the blank page or countering overthinking), and effort direction.

20.5.6 Work Effectiveness and Efficiency

Effort and performance are correlated (van Iddekinge et al., 2023), and according to *smart loafing*, enhancing work efficiency by relying on the technical system is a key concern (Stieglitz et al. 2022).

Time efficiency (↑) Interviewees reported improvements in both effectiveness (tackling harder tasks, I9) and efficiency (e.g., I1, I3-I4, I6, I9, I12-I15, I19-I21), with the emphasis on faster task completion. Reported subjective efficiency gains ranged between 60 and 90% in some cases (e.g., I6, I14) with one participant noting: “If I didn't have ChatGPT, I'm not sure if I would have actually managed the workload that I'm managing now” (e.g., I14). Interviewees realized these gains by GenAI helping them to prevent them from reading

irrelevant text, speeding up routine work (e.g., I4, I14-I15), researching topics difficult to google (I9, I12, I19), or composing documents by copying large amounts of GenAI output. However, efficiency gains seem to be much smaller for more conceptual tasks, one interviewee estimated around 3–5% improvement (I5), as much of the text is too generic and requires dedicated research and enrichment with expert knowledge. Some interviewees also drew comparisons with humans (e.g., working students), explaining that using GenAI was probably more cost-efficient (I10), the communication less ambiguous and the work results more neutral (I6).

Quality (?) The interviewees have mixed perceptions on how quality is affected by using GenAI. Some reported instances of quality improvement both in form and content, e.g., GenAI phrasing contents more eloquently (e.g., I9, I17), or cases in which GenAI proposed additions to their (conceptual) work results they had not thought about, but which were received positively by them or others (e.g., I6, I11). Others mentioned that the quality expectations and consequently the quality remained unchanged. Some interviewees felt that quality was slightly lower compared to working alone (e.g., I2) for certain tasks (e.g., I15). However, one interviewee (I2) mentioned the Pareto principle explaining that for many work settings, even 80% quality solutions are sufficient, and particularly given the saved time, a reduction in quality can be acceptable.

20.5.7 Cognitive Resources

According to smart loafing, a key idea of relying on technical systems is to maintain cognitive resources when completing a task (Stieglitz et al., 2022). While not frequently mentioned explicitly as a reasoning for using GenAI, even when asked about it directly, interviewees shared some reflections related to cognitive resources.

Relief through GenAI (↓) Most interviewees felt using GenAI to be a relief. They used it, when they had insufficient mental capacity to work on a task (e.g., I5, I19) or they felt lazy (e.g., I17). It helped freeing them from routine or operational work, providing them with more time to perform other tasks like strategic, creative or conceptual work (e.g., I9, I11-I13). Most participants explained that, if they did not perceive using the system as a relief, they would not use it and that doing the same work alone (as compared with GenAI) would require more “brain work” (I10, I19, I20), indicating at a potential reduction of cognitive load.

Stable external expectations (→) However, participants explained that the fact that they use GenAI does not change the stakeholders' expectations, or, as an interviewee (I2) put it: “The pressure comes more of the deadline or stakeholders and they don't really care how the product came about [...], so they don't care whether I spent 5 h on it or 100 h, the quality has to be right. And that goes hand in hand with my feeling that my pressure or stress [...] hasn't really changed. Just because you have a model like this, you can work faster, but the expectations for the result remain the same.” While GenAI can support task aspects, it also produces output requiring verification, creating additional cognitive load (e.g., I17, I19).

Because result expectations still must be met, and – as before – the worker is held accountable for the results, the cognitive load or stress remains the same (e.g., I2, I6), irrespective of how the results are produced. This point is closely connected to the external responsibility mentioned in the section on perceived responsibility.

Continuous learning (↑) Several participants reported putting in additional effort to learn to use the tools effectively (also see section on expectancy). Some interviewees mentioned aspects which do or could contribute to an increase in cognitive load, because they not only have to learn to use GenAI in the beginning but to continuously learn how to use it (e.g., I11, I17). Additionally, they have to test new versions, models, or modules to assess their potential for certain (business) tasks and business functions (e.g., I5, I9-I10, I16), which some described as exhausting (e.g., I5, I16). Furthermore, several interviewees shared that GenAI's output is not always satisfactory to them. While not explicitly described as a burden, multiple interviewees report that when the output of GenAI is not satisfactory, they believe that they have made a mistake (e.g., I9, I11, I8) and that the system is “cumbersome” (e.g. I14). This continuous learning can be perceived as additional (cognitive) effort, besides working on the task. One participant (I4) how the continuous learning affected her as follows “More and more [tools] are being thrown onto the market and once a week, someone comes around the corner with a different idea that you could try out [...] I wonder if someone else can try it out and tell me how well it works, because I find it exhausting when you try to understand a new tool every week and have to test out where the limitations are.”

Additional task (?) Lastly, different layers need to be considered regarding relieving professionals through GenAI: task- vs. job-level. Many interviewees state that they feel relieved and save time due to using GenAI for the tasks they perform. However, when asked about what they do with this saved time, they usually do not invest it into enhancing the quality, nor do they use it as free time, but rather they use it to tackle additional new tasks, or as one interviewee put it: “I have so much on my plate [...] there's always something else on the To-Do list filling up the 30 min of saved time” (I12). Thus, while relief might occur on the individual task level, on the overall job level, there might be a “fill-up” effect. Only two interviewees reported that using GenAI gives them some time off in between tasks or for having a coffee (I10-I11).

20.6 Discussion

20.6.1 GenAI's Effect on Human Behavior and Relationships

Our findings reveal three linked insights into how working with GenAI can change human behavior, specifically, how GenAI can reshape human motivation (1) and how humans invest effort (2). Additionally, the results indicate that human relationships in work contexts are related to working with GenAI (3).

Reshaped motivational antecedents Expectancy theory (Vroom, 1964) and the Collective Effort Model (Karau & Williams, 1993) predict that when workers believe that their effort will not meaningfully enhance performance, e.g., when working with a highly capable partner, workers will withhold effort. Prior studies extended this logic to AI predicting effort reductions (Liu et al., 2023), akin to social loafing in human groups, concluding that “[social loafing] occurs in virtual collaboration with [virtual assistance]” (Stieglitz et al., 2022, p. 759).

In our study, participants working with GenAI feel their effort to be dispensable for some (sub-)tasks. Overall, however, they report that they can still improve performance either by working on the work product directly or indirectly through prompting (*expectancy*), sometimes even feeling enabled. They perceive their involvement to be essential to achieve desired outcomes (*instrumentality*), stressing their process administration role as well as the necessity to provide context-specific knowledge and to shape and communicate the work product to meet stakeholder expectations. This contrasts with expectations of eroded necessity (Liu et al., 2023).

Social loafing in human groups frequently stems from diffused responsibility, when contributions are integrated into a single work result and are not identifiable. Earlier work suggests that such responsibility diffusion may not be limited to human groups (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022), but that humans may similarly assign responsibility to AI, stating that “the responsibility of solving a task in virtual collaboration is likely to be attributed to a [virtual assistant]” (Stieglitz et al., 2022, p. 759). Stieglitz et al. (2022, p. 759) concluded that for “virtual collaboration with [virtual assistants]” their study “indicates that knowledge from human-to-human collaboration in terms of [social loafing] is transferable”. Our data, however, suggests that social loafing-related knowledge from human groups is not necessarily transferable to professional human-GenAI settings when it comes to the identifiability of contributions and diffused responsibility. Potential explanations might be found in the specific characteristics of GenAI and its contributions, as well as in the fact that the tasks in our study are embedded in real work contexts in relation to other tasks, fellow humans and organizational expectations.

Literature on social loafing shows that when individual inputs merge into a single group product and cannot be identified, *responsibility diffuses* and effort tends to be reduced (i.e., social loafing occurs, Karau & Williams, 1993; Latané et al., 1979). If, on the other hand, an individual’s contribution is (a) *identifiable* or *known*, and (b) there is a *standard* against which the contribution can be compared, such loafing behavior may be counteracted (Karau & Williams, 1993). In our study, many interviewees report that recipients may struggle to separate their contribution from the GenAI’s (i.e., contributions are not *identifiable*) as GenAI outputs often resemble human contributions. Due to the iterative nature of working with GenAI, with constant re-integration of own and genAI contributions, individual contributions might be even more difficult to discern, as was also suggested in prior literature on human-AI teaming (Rix & Hess, 2022). Additionally, the interviewees do not always

disclose having used GenAI (i.e., contributions are not *known*). According to classical social loafing literature, this might lead to diffused responsibility. However, most interviewees report to still feel fully responsible. They cite both a professional ethos and explicit organizational expectations (i.e., a *standard*): the final work result carries their name, regardless of how much help the GenAI provided. This finding is contrary to prior studies, which showed that humans assigned responsibility to AI (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022).

Potential explanations for the divergence of our findings compared to predictions and findings on human-AI dyads (Henkenjohann & Trenz, 2024; Liu et al., 2023; Stieglitz et al., 2022) may be the contextual factors. First, the specifics of GenAI differ from the technical systems used in prior studies. Producing high-quality output requires iterative prompting: users must steer, enrich, and repeatedly verify responses that are probabilistic and potentially hallucinated. This sustained, hands-on work and the need to integrate results into the organization keeps human's expectancy and instrumentality high. Second, unlike brief lab tasks, our participants operate in authentic professional settings, where deliverables bear their name, and serve real stakeholders; established accountability norms anchor responsibility to humans, counteracting diffusion observed in controlled experiments (Henkenjohann & Trenz, 2024; Stieglitz et al., 2022; also see section "GenAI usage and relationships with others").

For now, GenAI appears to reshape rather than erode motivation with antecedents of effort remaining largely intact, even as workers incur new learning and verification burdens (see next section). Figure 2 synthesizes qualitative insights and provides a basis for testable propositions such as (exemplary selection):

Proposition 1 *The stronger a worker's belief that learning to steer GenAI improves its output, the higher their expectancy and the lower their intention to reduce effort for the task.*

Proposition 2 *The stronger a worker's belief that human process administration is indispensable when working with GenAI, the higher their instrumentality and the lower their effort-reduction intention for the task.*

Proposition 3 *If workers expect an unshared reward (i.e., full personal credit) when working with GenAI, the perceived valence (i.e., the desirability of the outcome) remains as high as for unaided work and they do not reduce their effort on the task.*

Proposition 4 *The stronger an organization's accountability norms, the less probable it is that work with GenAI will be felt to reduce responsibility.*

Suggested antecedents and relationships in Figure 2 also highlight conditions under which effort reduction could emerge, e.g., if future systems (a) consistently outperform humans even in expert tasks – lowering *expectancy*; (b) autonomously orchestrate workflows and interface with stakeholders – lowering *instrumentality*; (c) lead colleagues to withhold credit for outputs that can be generated by GenAI – lowering *valence*; or (d) are perceived as genuine teammates (Seeber et al. 2020), enabling *responsibility* diffusion. Tracking such technological and social shifts will be essential for anticipating when effort investment might slide toward loafing-like behavior.

Effort investment beyond effort reduction Prior literature on human–AI work has used a social-loafing lens (Liu et al. 2023; Stieglitz et al. 2022), predicting – or experimentally showing – lower human effort when working with AI, and has recently refined that idea into smart loafing: a deliberate effort reduction to gain efficiency and conserve cognitive resources (Stieglitz et al. 2022). Our interviews partially confirm that pattern on a (sub)task level: knowledge workers often reduce effort intensity for (sub-)tasks which GenAI handles well. However, participants also report that GenAI helps them to overcome hurdles in longer work processes – such as low energy or the “blank page” problem – thus maintaining momentum in more cognitively demanding tasks and increasing their *persistence*. We therefore propose:

Proposition 5 *Working with GenAI increases effort persistence.*

Perhaps more importantly, many report to deliberately shifting their effort direction by selectively using GenAI for certain (sub-)tasks (e.g., initial draft generation) and investing the freed-up time in higher-level thinking for the same or an additional task; hence, we propose:

Proposition 6 *When working with GenAI in a job context, reduced effort intensity for certain sub-tasks is – at least partially – offset by redirected effort into other sub-tasks or additional tasks.*

This seems to result in a “fill-up” effect on job-level, where time saved on one task is used to perform additional tasks. Although it is unclear whether working on these additional tasks adds (similar) value as compared to working more intensively on the original task, this behavior is a deliberate, context-dependent approach to allocating effort, rather than a simple, across-the-board reduction in effort as the term “loafing” might suggest. Extant literature has investigated effort reduction behavior in experiments in which participants had to perform one task several times in a row (e.g., Stieglitz et al. 2022); future studies should consider investigating settings containing *multiple, different* tasks to uncover the dynamic effort redirection in GenAI work (cf. Dell’Acqua et al. 2023).

A further layer of effort emerges when considering the uncertainty GenAI introduces to the effort-outcome link. When people work alone, the work result is largely dependent on their effort; with GenAI, it depends on two additional contingencies. First, the GenAI must – in principle – be capable of producing the content for the desired outcome. Second, the user must be able to elicit that content through effective prompting. To reduce uncertainty, respondents invest effort up front and continuously: they build prompt libraries, read blog posts, follow tutorials, and iteratively work with GenAI. These activities often pay off but are sometimes a time sink – time lost to hallucinations, style mismatches, or fruitless prompt tweaks. This is consistent with findings from Dell’Acqua et al. (2023), where a dedicated GenAI training (contrary to expectations) led to a performance reduction. This parallels the contingencies highlighted in the Collective Effort Model (Karau and Williams 1993): just as performance in human groups depends on others, performance in human–GenAI dyads depends on an uncertain technical “partner”. A purely outcome-based

perspective risks failing to acknowledge the specifics of working with quickly evolving, probabilistic GenAI systems, in which (up-front and continuously) invested effort does not guarantee improved performance (i.e., additional effort does not necessarily yield better results). When workers let GenAI draft texts, the visible drop in keystrokes may coincide with a redirection of effort to strategic analysis (i.e., intensity reduction at task level, but effort redirection at job level).

Overall, we argue for shifting the focus from asking whether humans reduce effort to examining how and why they invest it. A broader lens therefore needs to be multi-dimensional (different effort dimensions), multi-level (task and job), and longitudinal (capturing up-front learning and ongoing prompt-tuning). It should acknowledge the contingencies introduced by GenAI, where more effort does not always yield better results. Calling such behavior loafing – defined by Merriam-Webster (2024) as “to spend time in idleness” – risks mis-classifying effort re-allocation as idleness. A vocabulary centered on effort investment might better capture knowledge workers’ GenAI usage behavior.

GenAI use and relationship with others A difference of our compared to prior experimental studies is that our study is embedded in work contexts with real relationships to the recipients of the results produced with GenAI. Though not at the core of our study, our data suggests that human relationships might influence and be influenced by the work with GenAI beyond the impact of GenAI on the individual human (behavior). For example, this could occur with regard to *contribution identifiability* and *GenAI-use disclosure*, *unshared rewards*, or *satisfactory GenAI output*.

As discussed in the section on identifiability, despite human and GenAI contributions being closely integrated, workers still feel responsible for these work results due to organizational norms and stakeholder expectations. Hence, the decision to disclose or not disclose GenAI usage seems tactical rather than a way to enable social loafing. Interviewees decide on disclosure or non-disclosure based on factors such as the relationship with the recipient, or the importance or stage of the work result. Disclosure may help to manage expectations or spread good practice; non-disclosure may feel appropriate when GenAI’s role was minimal (akin to a Google search). In short, selective disclosure is tied to a variety of contextual motives and is not necessarily a means of evading effort. This relational pattern differs from human-only groups: even with low identifiability, responsibility can remain firmly anchored to the human so long as organizational norms or stakeholder relationships assign it there.

Relational dynamics extend to rewards and refining GenAI outputs. Interactions with colleagues influence motivational drivers (and consequently effort), e.g., if the praise or *reward* a human receives is made dependent upon whether and to which degree a work result was produced with a GenAI, this may affect the *valence* (i.e., the desirability of the outcome) of the individual sharing the result. Additionally, stakeholders’ quality expectations may affect whether GenAI output is considered *satisfactory* or needs further *refinement*,

affecting the person working with GenAI. Inversely, when sharing work results without sufficient quality assurance, workers may transfer workload to others or introduce risks. These factors have important implications for managing GenAI usage in organizations. Future research may employ organizational case studies – examining sender–recipient pairs or networked interactions – to reveal how social context mediates GenAI use and effort allocation.

20.6.2 Contribution to Theory

First, our study details how motivational antecedents are reshaped when knowledge workers use GenAI. Laboratory studies have predicted (Liu et al. 2023) or demonstrated effort reduction under certain controlled conditions and concluded that social loafing “occurs” when working with AI (Stieglitz et al. 2022, p. 759). Our study nuances this claim for a professional work context, finding that expectancy, instrumentality, and responsibility can stay high, while valence can remain stable, so a behavior that resembles loafing is not inevitable. Figure 2 offers a detailed view on how GenAI can increase, decrease, or redirect these antecedents based on its capabilities and usage. The model offers a testable map for future experiments and can serve as an aid for managers seeking to understand when and how GenAI is likely to erode or amplify motivation.

Focusing on a specific antecedent of effort reduction – diffused responsibility – our findings (1) reveal a potential boundary condition that complicates the classic link between identifiability and perceived responsibility, and (2) challenge prior literature that observed a shift of responsibility toward AI. In human groups, low identifiability of inputs allows responsibility to diffuse, setting the stage for loafing (Karau and Williams 1993). When working with GenAI, iterative prompting can weave GenAI and human contributions into a single result, i.e., the human contribution may have low identifiability – yet professionals continue to feel responsible. We suggest that this finding which is contrary to prior studies (cf. Henkenjohann and Trenz 2024; Stieglitz et al. 2022) may be due to organizational norms: low identifiability does not necessarily undermine responsibility if the “other member” is a GenAI and the organizational context keeps credit and blame with the individual. Future research should treat identifiability and transfer of responsibility as contingent on local accountability norms; only when norms allow diffusion should indistinguishability translate into lower responsibility and effort. Selective disclosure of GenAI use therefore does not imply an evaded effort, but it occurs as a situational signal.

Secondly, we reframe the recent discourse beyond effort reduction on the level of a specific task, towards a (broader) *effort investment lens*. Prior, lab-based studies measured effort (or loafing-like behavior) on a focal task and linked reduced effort to saved cognitive resources (Stieglitz et al. 2022). Our data confirms such reduced effort on a (sub-)task level but shows that workers may redirect time to – according to their perspective – higher-value (sub-)tasks or additional tasks, potentially resulting in a job-level fill-up effect. Moreover, participants invest considerable effort into improving their skills. Understanding such patterns requires

tracking intensity, persistence, and direction of effort across tasks and over time, and factoring in the probabilistic, evolving nature of GenAI. Rather than abandoning loafing theory, we therefore suggest a more comprehensive *effort investment* perspective, shifting the core question from “Does effort become less?” to “Where is effort invested, and why?” – a question that calls for longitudinal, multi-dimensional analyses able to distinguish idleness from strategic re-allocation. While controlled experimental studies are important to measure effort on specific tasks, our qualitative study sheds light on the broader interrelations in work settings. This study provides indication that an observed reduction of effort on a focal task might have different reasons, including not only behavior analogous to loafing, but also investment in future skills or shifts to other tasks. Additionally, the terminology of *investment* reflects the common belief among our participants that the effort spent to learn how to effectively work with GenAI will pay off, while acknowledging the uncertainty of returns related to probabilistic, quickly evolving GenAI systems.

20.6.3 Implications for Practice

Our findings highlight that working with GenAI in professional settings can reshape – rather than merely reduce – human motivation and effort. However, whether workers can leverage GenAI’s benefits depends on contextual factors: policies, expectations, and relationships must adapt to reflect the specifics of working with GenAI. These specifics include the quickly evolving capabilities of GenAI and the uncertainty surrounding working with GenAI, where more effort does not always yield better results. Managers play a critical role in shaping these conditions, e.g., by designing supportive environments, clarifying responsibilities, and enabling effective use of GenAI.

Encourage exploration, but acknowledge learning costs and uncertainty Aligning with our broader conceptualization of effort investment, managers should recognize that up-front and continuous learning in effectively steering GenAI and integrating it into the workflow is critical for leveraging GenAI effectively. Encouraging employees to experiment and allowing for dips in efficiency may pay off. Viewing inefficiencies as “wasted effort” does not acknowledge the contingencies associated with using ever changing, probabilistic GenAI for knowledge work, where additional effort may not always translate into improved performance. A supportive environment – where employees feel safe to iterate and even fail – can ultimately support the development of competencies for GenAI usage.

Clarify expectations and responsibilities Organizations seeking to develop policies for GenAI usage should carefully consider the new ways of working with GenAI. Knowledge workers adjust the effort they invest into tasks, how much they use GenAI, and how much effort they spend on verifying GenAI outputs dynamically, enabling them to quickly iterate on their path toward final work results. Static policies like “You are responsible for the quality of your results” may not adequately reflect the dynamic use of GenAI to quickly produce drafts results to share with colleagues that we have observed in our data; instead, it may prevent workers from using GenAI if they do not have the expertise or time to verify

outputs. A more flexible approach, e.g., asking workers to communicate the stage or quality of the shared result, may be more helpful, as it allows workers to benefit from the potential of GenAI while helping to prevent potential risks or causing additional work.

Situational disclosure Our findings suggest that decisions around disclosing GenAI usage are rarely about evading effort; instead, they serve as contextual signals that help to communicate quality levels, manage expectations, or advocate GenAI use within teams. Therefore, mandating universal disclosure of GenAI use may be counterproductive. It could inadvertently lower motivation (valence) if colleagues or managers assume that “the GenAI did all the work,” diminishing perceived individual contribution. It might also encourage “lazy draft” behavior where workers offload low-quality outputs onto others, assuming disclosure shifts the burden of review. Rather than blanket policies, organizations should consider implementing flexible, context-sensitive disclosure guidelines. These should support building a shared understanding of a work result’s quality, allowing workers to benefit from GenAI while maintaining responsibility in collaborative settings.

20.6.4 Limitations and Further Research

This study employed a qualitative, interview-based approach, which poses certain challenges in assessing changes in effort. First, social desirability bias (Döring and Bortz 2016) may have led participants to only mention selected interactions or downplay reductions of effort, especially given the negative connotations of “loafing.” To mitigate this, we promised anonymity to the interviewees, built rapport prior to addressing sensitive topics, explained that we seek to understand their experience, and asked indirect questions, e.g., also inquiring about antecedents and considering behaviors, even if participants did not label them as “reduced effort” directly. Second, participants might unconsciously reduce effort, truthfully reporting their experiences but remaining unaware of underlying effort shifts. We tried to counter this and to uncover effort changes that interviewees had not been aware of during the situation (i.e., unconscious effort changes) by asking them to walk us through specific tasks completed with GenAI and reflecting on their behavior to better understand their decision-making and potential “preexisting effort scripts” (Karau and Williams 1993, p. 685). Nonetheless, potential unconscious effort changes may have been missed by our approach; future research might mitigate this through triangulation, e.g., using trace data or in-situ observations. Third, to reduce subjectivity during the analysis, we followed a directed content analysis approach with predefined categories (based on our conceptual model) and incorporated representative quotes; nonetheless, potential biases persist. Lastly, for our directed content analysis, we derived a conceptual model integrating previously discussed antecedents and consequences for effort investment when working with AI. However, literature on expectancy theory and social loafing is vast and we encourage exploring additional factors.

We conducted interviews with a relatively small number of knowledge workers spanning various professional domains, though nearly half were researchers. While we reached saturation – no new insights emerged from subsequent interviews – the overrepresentation of researchers may have caused us to conclude saturation prematurely, and interviewing additional participants in other roles or industries might uncover different or more nuanced behaviors. Moreover, most interviewees primarily used ChatGPT or Microsoft Copilot, which could influence our findings. However, as these systems are widely used, they might be representative of professionals' experiences. While participants varied widely in their prevalent tasks, we drew on established theories (expectancy theory, social loafing) that have been validated across diverse settings. The broad sampling with explicit selection criteria determined by the study goal aimed to capture general effort-related patterns rather than confining the study to one domain.

Our data reflects participants' perceptions at specific points in time, during which organizations and individuals are still adapting to GenAI. Future, more capable or agentic systems which are able to autonomously discover tasks and interact with stakeholders could further alter motivation or reduce perceived responsibility, making loafing-like behavior more likely. Additionally, changes in perception of GenAI-assisted work may shift valence or instrumentality. Our model (Figure 2) may serve as a starting point to (quantitatively) study the effects of these developments on human motivation and effort.

Moving forward and building on the “smart loafing” concept (Stieglitz et al. 2022), future research should consider *effort investment* when working with GenAI, also reflecting the additional contingencies that occur. Such work might examine longitudinal changes in effort distribution as people gain proficiency, capture real-time GenAI usage data, or investigate organizational practices (e.g., disclosure policies) that shape how humans balance their tasks with GenAI support. Additionally, investigating sociotechnical solutions for sharing GenAI-assisted outputs – such as standardized labeling for draft quality or embedded revision histories – could prevent miscommunications, manage stakeholder expectations, and mitigate potential additional effort or risks. Embracing these avenues will allow future IS studies to deepen our understanding of how knowledge workers meaningfully integrate rapidly evolving GenAI into their work, thereby preventing some of the aforementioned challenges.

20.7 Conclusion

GenAI is increasingly woven into professional knowledge work and contrary to fears of widespread effort reduction, our study suggest that knowledge workers deliberately manage their effort: while they rely on GenAI for certain subtasks, they also invest effort in learning to prompt, verify, and integrate GenAI outputs. Simply transferring the social loafing lens from purely human groups to human–GenAI dyads risks overlooking this dynamic

interplay – where up-front learning, strategic reallocation of effort, and an uncertain relationship between worker input and GenAI-generated outputs all come into play. We thus propose a more nuanced perspective, highlighting ways of how GenAI might reshape effort investment both directly and indirectly, and offer initial practical implications for managers and organizations. However, as GenAI capabilities evolve, further research is needed to refine our understanding of how professionals can optimize their effort investment – maintaining responsibility, enhancing performance, and upholding motivation – in an era of rapidly advancing GenAI.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1007/s12599-025-00960-4>.

Acknowledgements

This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

Funding

Open Access funding enabled and organized by Projekt DEAL.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

20.8 References

- Cymek, D. H. (2018). Redundant Automation Monitoring: Four Eyes Don't See More Than Two, if Everyone Turns a Blind Eye. *Human Factors*, 60(7), 902–921. <https://doi.org/10.1177/0018720818781192>
- Cymek, D. H., & Manzey, D. (2022). Sequential human redundancy: Can social loafing diminish the safety of double checks? *Journal of Experimental Psychology. Applied*, 28(4), 931–945. <https://doi.org/10.1037/xap0000439>
- Cymek, D. H., Truckenbrodt, A., & Onnasch, L. (2023). Lean back or lean in? Exploring social loafing in human-robot teams. *Frontiers in Robotics and AI*, 10, 1249252. <https://doi.org/10.3389/frobt.2023.1249252>
- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraye, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Working Paper*.
- Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. <https://doi.org/10.1007/s12599-019-00595-2>

- Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-642-41089-5>
- Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough? An Experiment with Data Saturation and Variability. *Field Methods*, 18(1), 59–82. <https://doi.org/10.1177/1525822X05279903>
- Haase, J., & Hanel, P. H. P. (2023). Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity. *Journal of Creativity*, 33(3). <https://doi.org/10.1016/j.yjoc.2023.100066>
- Hamza, M., Siemon, D., Akbar, M. A., & Rahman, T. (2024). Human-AI Collaboration in Software Engineering: Lessons Learned from a Hands-On Workshop. In *7th ACM/IEEE International Workshop on Software-intensive Business* (pp. 7–14). ACM. <https://doi.org/10.1145/3643690.3648236>
- Henkenjohann, R., & Trenz, M. (2024). Challenges in collaboration with Generative AI: Interaction patterns, outcome quality and perceived responsibility. In *ECIS 2024 Proceedings*, Paphos, Cyprus.
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.
- Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base. *Reuters*. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- Hüffmeier, J., Dietrich, H., & Hertel, G. (2013). Effort Intentions in Teams: Effects of Task Type and Teammate Performance. *Small Group Research*, 44(1), 62–88. <https://doi.org/10.1177/1046496412472242>
- Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4), 681–706. <https://doi.org/10.1037/0022-3514.65.4.681>
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *J. Pers. Soc. Psychol.*, 37(6), 822–832. <https://doi.org/10.1037/0022-3514.37.6.822>
- Liu, M., Ke, W., & Xu, D. J. (2023). Will Humans be Free-Riders? The Effects of Expectations for AI on Human-AI Team Performance. *PACIS 2023 Proceedings*.
- Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, 262–273. <https://doi.org/10.1016/j.jbusres.2020.07.045>
- Mandt, E. (1978). Managing the Knowledge Worker of the Future. *Personnel Journal*, 57(3), 138–143.
- Mayring, P. (2014). *Qualitative Content Analysis - Theoretical Foundation, Basic Procedures and Software Solution*.

- Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In *Proceedings of the 55th Hawaii International Conference on System Sciences*, Virtual Event/Maui, HI, USA. <http://hdl.handle.net/10125/79376>
- Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In *Proceedings of the 57th Hawaii International Conference on System Sciences*, Honolulu, HI, USA. <https://hdl.handle.net/10125/106425>
- Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In *Proceedings of the 45th International Conference on Information Systems (ICIS)*, Bangkok, Thailand. <https://aisel.aisnet.org/icis2024/aiinbus/aiinbus/7/>
- Memmert, L., Soroko, D., & Bittner, E. (2025). From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI. *Business & Information Systems Engineering*(67), 615–635. <https://doi.org/10.1007/s12599-025-00960-4>
- Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. *Journal of Information Systems Education*, 34(3), 279–292. <https://aisel.aisnet.org/jise/vol34/iss3/3>
- Meuser, M., & Nagel, U. (2009). Das Experteninterview — Konzeptionelle Grundlagen und Methodische Anlage. In S. Pickel, G. Pickel, H.-J. Lauth, & D. Jahn (Eds.), *Methoden der vergleichenden Politik- und Sozialwissenschaft* (pp. 465–479). VS Verlag für Sozialwissenschaften.
- Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Onnasch, L., & Panayotidis, T. (2020). Social Loafing with Robots – An Empirical Investigation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1), 97–101. <https://doi.org/10.1177/1071181320641026>
- Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, 10(2), 110–133. <https://doi.org/10.1287/isre.10.2.110>
- Pyöriä, P. (2005). The concept of knowledge work revisited. *Journal of Knowledge Management*, 9(3), 116–127. <https://doi.org/10.1108/13673270510602818>
- Retkowsky, J., Hafermalz, E., & Huysman, M. (2024). Managing a ChatGPT-empowered workforce: Understanding its affordances and side effects. *Business Horizons*, 67(5), 511–523. <https://doi.org/10.1016/j.bushor.2024.04.009>

- Rix, J. (2022). From tools to teammates: Conceptualizing humans' perception of machines as teammates with a systematic literature review. In *Proceedings of the 55th Hawaii International Conference on System Sciences*.
- Rix, J., & Hess, T. (2022). Hello, Mate! Insights from the Field on Leveraging Machine Teammates in Organizations. In *PACIS 2022 Proceedings*. <https://aisel.aisnet.org/pacis2022/95>
- Seeber, I., Bittner, E., Briggs, R. O., Vreede, T. de, Vreede, G.-J. de, Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, 57(2). <https://doi.org/10.1016/j.im.2019.103174>
- Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *48th HICSS Proceedings*.
- Stieglitz, S., Mirbabaie, M., Möllmann, N. R. J., & Rzycki, J. (2022). Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution. *Information Systems Frontiers*, 24(3), 745–770. <https://doi.org/10.1007/s10796-021-10201-0>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Timonen, H., & Paloheimo, K.-S. (2008). The emergence and diffusion of the concept of knowledge work. *Electronic Journal of Knowledge Management*, 6, 177–190.
- van Iddekinge, C. H., Arnold, J. D., Aguinis, H., Lang, J. W. B., & Lievens, F. (2023). Work Effort: A Conceptual and Meta-Analytic Review. *Journal of Management*, 49(1), 125–157. <https://doi.org/10.1177/01492063221087641>
- Vroom, V. H. (1964). *Work and motivation*. Wiley.

20.9 Appendix

20.9.1 Appendix A: Interview Guideline

Table A.1 Interview guidelines used in the study and formatted for clarity. The guideline was used in a semi-structured approach, where the questions provided a structure to the interviews but left room for participants to share their experiences, consistent with best practices in qualitative research.

ID	Questions
Introduction + agreement	
Q0	Do you agree that we record the interview, use a software for automatic transcription, and use the contents for anonymized publication?
Socio-demographics	
Q1	What is your current job title?
Q2	How many years of work experience do you have?
Experience with generative language models	

ID	Questions
Q3	What kind of GLMs do you use and which versions?
Q4	How long have you used GLMs for?
Q5	How frequently do you use it?
Q6	For what kind of tasks do you use GLMs? Can you give examples of tasks where you have worked alone with the GLM (i.e., 1 human: 1 GLM setting)?
Q7	Are there examples of tasks where the results are given to third parties? If yes, let's focus on them moving forward.
Approaches/techniques/strategies for working with the GLM	
Q8	When working with GLMs: How do you integrate GLMs into your workflow? Have you developed any routines or approaches when you want to solve a problem or work on a task?
Q9	How do you choose which task to do on your own and which one to complete with a GLM?
Q10	How do you determine when to stop working with the GLM on the task? Have you ever had the feeling that working with a GLM was a waste of time? If yes, what did you do about it? How do you prevent that from happening?
Q11	Do you, and if so, how do you determine and ensure the quality of results? Do you use the GLM's outputs directly? How has the result quality changed (if it has)?
Q12	How do you use the output from the GLM for your work? Is it communication in constant change or one-time communication? Why is that?
Q13	Have you discovered approaches that do not seem to work for you? Which and why? Have you evolved or adjusted your approach?
Effort	
Q14	How do you see your role when working with ChatGPT or other GLMs?
Q15	Do the work results feel in any way different when achieved with GLM compared to when developed alone (or with Google)? Why/why not?
Q16	Does ChatGPT replace parts of your work? If so, does this change how you think about your work and work products?
Q17	Could you use the output of ChatGPT directly, without modification?
Q18	Do you communicate that you have worked with the GLM to the person you are providing the results to?
Q19	Do you think that the recipients of the work results can distinguish your and the GLMs' contribution?
Q20	Has feedback on your work results ever been directly linked to your work with the GLM (positive or negative)?
Q21	How does working with the GLM affect you and the effort you have to invest?
Q22	How does the use of GLMs affect your stress or mental strain at work?
Q23	Does the GLM contribute something that you couldn't have solved on your own?
Q24	And vice versa: Do you contribute something that the GLM couldn't have delivered?
Q25	Do you see any negative consequences of working with the GLM?
Q26	How does working with the GLM affect your work efficiency?
Q27	If you save time by working with GLMs, how do you use this saved time?
Q28	Does it require significant effort to work well with GLMs?
Q29	When you are working on a task, would you rather delegate it to another human or do it with ChatGPT?
Q30	When you work with a GLM, does this affect how hard you work on the task?
Q31	When solving problems, sometimes one runs into obstacles. Has that happened when you worked with GLMs and how do you see the role of GLMs in this regard?
Q32	How well do you feel you can control the GLM to do the things you want it to do?
Wrap-up	
Q33	Is there anything else you would like to share?

20.9.2 Appendix B: Coding Scheme

Table B.1 Coding scheme with categories, subcategories, and examples

Category	Subcategory	Example
Expectancy	Own contribution	“The results [produced by the GenAI] are semi usable if you don't put your own know-how into it [...]. What we do requires a certain creative approach, which the systems currently do not yet provide” (I4).
	(Learning to) steer the GenAI	“Especially at the beginning, as I said, I looked a lot at blog articles and things like that on how you can, may, should optimize your workflows with [GenAI]; what kind of prompt is suitable or whatever. But now I think I know that I at least have a set of things that work quite well for me.” (I5)
	Enablement via GenAI	“At the fingertips, I have so much power now that I can use!” (I12)
	Perceived time sink	“I spend so much time now essentially arguing with GPT about a solution that I'm not happy with. And this was really not fruitful in any way and kind of a waste of my time... Like one out of 10 times, I think this happens.” (I14)
	Perceived redundancy	“I sometimes do not find the right word, and [GenAI] makes it more eloquent” (I10)
Instrumentality	Process administration	“I think I'm good at conceptualizing and planning execution steps. And it's kind of like I have now an assistant. I can make this pretty elaborate and good plan and I can give it to someone to execute it and guide them along the way, which I feel like I'm very good at.” (I14)
	Main intellectual contribution	“[the task] it is too complex most of the time; ChatGPT just does not have the required background knowledge” (I8)
	GenAI output refinement	“In my opinion, we will never be spared the step of saying that a human who knows what they're doing needs to look at [the GenAI output] again to say whether it fits or not” (I3)
	Satisfactory GenAI output	“I think [the GenAI output] is of a very high quality. I think if I had designed [the work product] myself without the help of the AI, it would probably be 10 or 20 percent better, but usually 80% is enough in the work context” (I2)
Valence	Unshared reward	“Where [my manager] then joked and said ‘Yes, [name of the employee], why do we still need you at all then?’” (I10)
	Additional reward	“We had worked out various points as part of a strategy, and then we were asked, ‘Oh, I think that's particularly well written’, and that was exactly the point we didn't write ourselves.” (I3)
	Changed achievement feeling	“It feels a bit like someone else is solving the problem for you” (I9)
Perceived responsibility	Contribution identifiability	“I don't think you can actually distinguish it because there is not a visible ChatGPT contribution” (I21)
	GenAI-use disclosure	“With the people I work closely with; I usually make (ChatGPT use) transparent” (I6)
	External accountability	“Nobody cares about the way the results are produced – it is perceived as coming from me” (I2)
	Felt responsibility	“I'm still proud. And I own the idea of the whole thing. I will always also stand up for any mistakes that happen. Because I feel like, I'm putting these thing out, it's my responsibility” (I14)
Work effectiveness & efficiency	Time efficiency	“So the time [without GenAI] would have been significantly longer, [...], [with the GenAI] I needed maybe 1/8 of the time compared to as if I hadn't had GenAI available”
	Quality	“I don't think it has gotten better [...]” (I21)

Category	Subcategory	Example
Cognitive re-sources	Relief through GLM	"I do feel a little bit more relieved when I'm very stressed and I'm working on a time sensitive task and then I can ask something ChatGPT and then save some time. So that provides me some relief." (I12)
	Stable external expectations	"The pressure comes more of the deadline or stakeholders and they don't really care how the product came about [...], so they don't care whether I spent 5 hours on it or 100 hours, the quality has to be right. And that goes hand in hand with my feeling that my pressure or stress [...] hasn't really changed. Just because you have a model like this, you can work faster, but the expectations for the result remain the same." (I2)
	Continuous learning	"More and more [tools] are being thrown onto the market and once a week, someone comes around the corner with a different idea that you could try out [...] I wonder if someone else can try it out and tell me how well it works, because I find it exhausting when you try to understand a new tool every week and have to test out where the limitations are." (I4)
	Additional tasks	"I have so much on my plate [...] there's always something else on the To-Do list filling up the 30 minutes of saved time" (I12)

Full publication list

The table below lists all publications co-authored by Lucas Memmert as of December 2025.

No.	Publication	Included in this dissertation
1	Oeste-Reiß, S., Bittner, E., Cvetkovic, I., Günther, A., Leimeister, J. M., Memmert, L., Ott, A., Sick, B., & Wolter, K. (2021). Hybride Wissensarbeit. <i>Informatik Spektrum</i> , 44(3), 148–152. https://doi.org/10.1007/s00287-021-01352-0	No
2	Memmert, L., & Bittner, E. (2022). Complex Problem Solving through Human-AI Collaboration: Literature Review on Research Contexts. In <i>Proceedings of the 55th Hawaii International Conference on System Sciences</i> , Virtual Event/Maui, HI, USA. http://hdl.handle.net/10125/79376	Yes
3	Memmert, L., & Tavanapour, N. (2023). Towards Human-AI Collaboration in Brainstorming: Empirical Insights into the Perception of Working with a Generative AI. In <i>Proceedings of the 31st European Conference on Information Systems (ECIS)</i> , Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/429	Yes
4	Memmert, L., Cvetkovic, I., & Bittner, E. (2023). Human-AI Collaboration in Conceptualizing Design Science Research Studies: Perceived Helpfulness of Generative Language Model's Suggestions. In <i>Proceedings of the 31st European Conference on Information Systems (ECIS)</i> , Kristiansand, Norway. https://aisel.aisnet.org/ecis2023_rp/405	Yes
5	Memmert, L., Tavanapour, N., & Bittner, E. (2023). Learning by Doing: Educators' Perspective on an Illustrative Tool for AI-Generated Scaffolding for Students in Conceptualizing Design Science Research Studies. <i>Journal of Information Systems Education</i> , 34(3), 279–292. https://aisel.aisnet.org/jise/vol34/iss3/3	Yes
6	Brackel-Schmidt, C. von, Kučević, E., Memmert, L., Tavanapour, N., Cvetkovic, I., Bittner, E., & Böhm, T. (2023). A User-centric Taxonomy for Conversational Generative Language Models. In <i>Proceedings of the 44th International Conference on Information Systems (ICIS)</i> . https://aisel.aisnet.org/icis2023/generalis/generalis/1	No
7	Kučević, E., Brackel-Schmidt, C. von, Lewandowski, T., Leible, S., Memmert, L., & Böhm, T. (2024). The Prompt-a-thon: Designing a Format for Value Co-Creation with Generative AI for Research and Practice. In <i>Proceedings of the 57th Hawaii International Conference on System Sciences</i> , Honolulu, HI, USA. https://hdl.handle.net/10125/106577	No
8	Memmert, L. (2024). Brainstorming with a Generative Language Model: Understanding Performance Through Brainstorming Group Effects. In <i>Proceedings of the 32nd European Conference on Information (ECIS)</i> , Paphos, Cyprus. https://aisel.aisnet.org/ecis2024/track06_humanaicollab/track06_humanaicollab/1	Yes
9	Memmert, L., Mies, J., & Bittner, E. (2024). Brainstorming with a Generative Language Model: The Role of Creative Ability and Tool-Support for Brainstorming Performance. In <i>Proceedings of the 45th International Conference on Information Systems (ICIS)</i> , Bangkok, Thailand. https://aisel.aisnet.org/icis2024/aiinbus/aiinbus/7/	Yes
10	Memmert, L., & Bittner, E. (2024). Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration. In <i>Proceedings of the 57th Hawaii International Conference on System Sciences</i> , Honolulu, HI, USA. https://hdl.handle.net/10125/106425	Yes
11	Memmert, L., Cvetkovic, I., & Bittner, E. (2024). The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3. In <i>Proceedings of the 57th Hawaii International Conference on System Sciences</i> , Honolulu, HI, USA. https://hdl.handle.net/10125/107289	Yes
12	Memmert, L., Borchers, M., Plückhahn, J., & Bittner, E. (2025). Solving Coding Challenges Jointly with a Large Language Model: Understanding Student Journeys Through Bloom's Taxonomy. In <i>Proceedings of the 58th Hawaii International Conference on System Sciences (HICSS)</i> , Waikoloa Village, HI, USA. https://hdl.handle.net/10125/109714	Yes

No.	Publication	Included in this dissertation
13	Memmert, L., Soroko, D., & Bittner, E. (2025). From Effort Reduction to Effort Management: An Expectancy Theory Perspective on Professionals' Work Practices with Generative AI. <i>Business & Information Systems Engineering</i> (67), 615–635. https://doi.org/10.1007/s12599-025-00960-4	Yes
14	Memmert, L., Cvetkovic, I., Tavanapour, N., & Bittner, E. (2025). Brainstorming with a Generative Language Model: Effect of Exposure to AI Ideas on Brainstorming Performance and Cognitive Load. <i>Business & Information Systems Engineering</i> . Advance online publication. https://doi.org/10.1007/s12599-025-00974-y	Yes
15	Vreede, T. de, & Memmert, L. (2025). AI-Assisted Task Performance: The Moderating Role of Assistance Modality on Motivational Dynamics. In <i>Proceedings of the 46th International Conference on Information Systems (ICIS)</i> , Nashville, Tennessee, US. https://aisel.aisnet.org/icis2025/user_behav/user_behav/22	No

GLM usage while producing this dissertation

Aligned with the practices at the chair of Information Systems, Socio-Technical System Design (WISTS), and in agreement with the supervisor of this dissertation, Prof. Dr. Eva Bittner, this dissertation and many of the publications included were developed using GLMs in various capacities. In particular, I leveraged the opportunities offered by this new technology (particularly, OpenAI ChatGPT, Google NotebookLLM, and Grammarly) as discussed by Susarla et al. (2023), including but not limited to support the development of the research problem and design, to support the analysis of qualitative content, to support the development of the argument structure, and to support writing, including final phrasing and translation. GLMs were used while following the "suggestions for mitigating problems posed by generative AI" (Susarla et al., 2023, p. 406) and in accordance with the conference and journal guidelines of the respective publications. The authors are aware of the limitations of GLMs, retained full control over the research process, and are solely responsible and accountable for the results.

Reference

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, 34(2), 399–408. <https://doi.org/10.1287/isre.2023.ed.v34.n2>

Eidesstattliche Versicherung

Eidesstattliche Versicherung:

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

Affidavit:

I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated. If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.

Ort, den | City, date

Unterschrift | Signature

Identität eingereichter Versionen

Versicherung zur Identität der elektronischen und gedruckten Dissertationsschrift

Ich versichere, dass dieses gebundene Exemplar der Dissertation und das in elektronischer Form eingereichte Dissertationsexemplar (über den Docata-Upload) und das bei der Fakultät (zuständiges Studienbüro) zur Archivierung eingereichte gedruckte, gebundene Exemplar der Dissertationsschrift identisch sind.

I, the undersigned, declare that this bound copy of the dissertation and the dissertation submitted in electronic form (via the Docata upload) and the printed bound copy of the dissertation submitted to the faculty (responsible Academic Office) for archiving are identical.

Ort, den | City, date

Unterschrift | Signature