Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

DISSERTATION

Bridging Vision, Language, and Gaze for Trustworthy Foundation Models

Xintong Wang

Language Technology
Department of Informatics
Faculty of Mathematics, Informatics and Natural Sciences

Universität Hamburg
Hamburg, Germany

An der Universität Hamburg eingereichte monographische Dissertation
zur Erlangung des akademischen Grades

Doctor rerum naturalium (Dr. rer. nat.)

2026

Bridging Vision, Language, and Gaze for Trustworthy Foundation Models

Dissertation submitted by: Xintong Wang

Date of Submission: May 11, 2026
Date of Disputation: June 23, 2026

Supervisor: Prof. Dr. Chris Biemann, Universität Hamburg

Examination Commission:

Chair:	Prof. Dr. Sören Laue, Universität Hamburg
Deputy Chair:	Prof. Dr. Anne Lauscher, Universität Hamburg
Member:	Prof. Dr. Chris Biemann, Universität Hamburg
Member:	Assistant Professor Pinjia He, The Chinese University of Hong Kong, Shenzhen

Evaluators (Assessors):

1st Evaluator:	Prof. Dr. Chris Biemann, Universität Hamburg
2nd Evaluator:	Prof. Dr. Anne Lauscher, Universität Hamburg
3rd Evaluator:	Assistant Professor Pinjia He, The Chinese University of Hong Kong, Shenzhen

Universität Hamburg, Hamburg, Germany
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics

Language Technology

Affidavit

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.

I hereby declare upon oath that I have written the present dissertation independently and have not used any resources or aids other than those stated. Insofar as generative artificial intelligence (gAI) based electronic tools were used in preparing this dissertation, I affirm that my own contribution was paramount and that a complete documentation of all tools used is available in accordance with the principles of Good Scientific Practice. I bear responsibility for any erroneous or distorted content, faulty references, violations of data protection or copyright law, or plagiarism that may have been generated by gAI.

Date

Signature
(Xintong Wang)

To the people of the world suffering from war, poverty, and hunger.

May your pain never be forgotten.

Acknowledgements

This dissertation was completed during a period marked by uncertainty and distance, as it began in the early years of the global pandemic. Starting doctoral research under such conditions was challenging both academically and in adapting to new circumstances, and it required trust and patience from those involved. I am very grateful to my supervisor, Prof. Chris Biemann, for offering me the opportunity to pursue this path and for supporting me with confidence long before tangible results emerged.

Throughout the doctoral journey, Chris provided not only guidance but also an environment of intellectual openness and encouragement. In the early stages, when progress felt slow and adapting to a new academic setting was demanding, his steady reassurance helped me focus on developing my own research direction rather than comparing myself to external timelines. He consistently recognized effort, curiosity, and persistence alongside outcomes. Over time, this mentorship transformed uncertainty into motivation, and milestones that once seemed distant gradually became achievable. I am particularly thankful for the many opportunities he provided—to present work, mentor students, engage in collaborative discussions, and gain insight into how research communities and projects are built. These experiences shaped not only this dissertation but also my understanding of academic life.

I would also like to express my sincere appreciation to Prof. Xingshan Li at the Institute of Psychology, Chinese Academy of Sciences, and to Prof. Markus Hoffmann at the University of Wuppertal. Their perspectives opened important interdisciplinary directions during my doctoral studies. Exposure to cognitive science and discussions on human behavior, perception, and memory significantly influenced how I approached questions about language models and interpretability, broadening my view of research beyond disciplinary boundaries.

My gratitude extends to the students and collaborators who contributed to this research through their enthusiasm and willingness to explore new ideas together. Working with Jingheng Pan, Daryna Dementieva, and Alexander Panchenko turned many initial concepts into concrete results and made the research process a shared intellectual endeavor. I am likewise thankful to Dr. Liang Ding and Dr. Longyue Wang for their collaboration, insightful discussions, and support that enabled the practical realization of several research ideas. I would also like to acknowledge my colleagues in the Language Technology group, especially Hans, Fynn, Steffen, and Martin, for creating a collegial environment in which discussions, questions, and everyday exchanges continuously enriched the research experience.

On a personal level, I am deeply grateful to my parents. Their unwavering trust and encouragement provided the foundation that allowed me to pursue this journey with confidence. Their constant care and belief made it possible to remain focused during both demanding and uncertain periods of doctoral life.

But the truly marvelous, magnificent, and extraordinary sights of the world are always found in dangerous and remote places where few men venture; thus, only those with resolution can reach them.

而世之奇伟、瑰怪、非常之观，常在于险远，
而人之所罕至焉，故非有志者不能至也。

— Wang Anshi, Song Dynasty, 1054

Abstract

Foundation models have reshaped the development of artificial intelligence (AI) by introducing a paradigm in which a single pretrained system can generalize across a wide spectrum of language and multimodal tasks. Built upon large-scale data and representation learning, these models reduce the need for task-specific engineering and enable reusable semantic and perceptual knowledge. This shift from specialized models to general-purpose learning architectures has significantly expanded the scope and applicability of AI technologies. As foundation models transition from controlled benchmarks to real-world deployment, new demands arise that extend beyond raw capability. Systems are increasingly expected to behave in ways that are reliable, interpretable, and aligned with human expectations, particularly in settings involving multimodal reasoning and interaction. These requirements expose limitations that are not visible in traditional performance evaluations and motivate a shift in research focus from scaling performance to ensuring trustworthy behavior.

In this dissertation, trustworthiness is analyzed as a property that emerges from the structural coupling of groundedness, alignment stability, faithfulness, and controllability. These dimensions correspond to interdependent stages of the trustworthiness pipeline, encompassing how multimodal signals anchor meaning, how pretrained representations are adapted, how generative processes reconcile internal knowledge with external evidence, and how model behavior can be guided in transparent ways. When these stages are treated in isolation, characteristic failure modes arise, including context-insensitive grounding, instability under adaptation, hallucinated outputs, and safety interventions that disrupt communicative intent.

The dissertation therefore investigates trustworthiness through coordinated interventions at different interfaces of the modeling pipeline. It develops methods that strengthen context-sensitive multimodal grounding while preserving representational structure, examines inference-time mechanisms that regulate the interaction between prior knowledge and conditioning signals, and introduces cognitively informed analyses to identify interpretable loci for efficient behavioral steering. Rather than addressing isolated symptoms, these contributions target complementary sources of unreliability across data construction, representation maintenance, and generation dynamics.

Overall, the findings indicate that trustworthy foundation modeling must be engineered as a lifecycle property rather than achieved through post hoc alignment alone. Reliability arises from the deliberate coordination of grounding, adaptation, inference, and control, suggesting a pathway toward foundation models whose general capabilities are matched by predictability, transparency, and human-centered usability.

Zusammenfassung

Grundlagenmodelle (Foundation Models) haben die Entwicklung künstlicher Intelligenz grundlegend geprägt, indem sie ein Paradigma etabliert haben, in dem ein einzelnes vortrainiertes System auf ein breites Spektrum sprachlicher und multimodaler Aufgaben generalisieren kann. Auf der Basis großskaliger Daten und repräsentationsbasierter Lernverfahren ermöglichen diese Modelle eine Wiederverwendung semantischer und perzeptueller Wissensbestände über unterschiedliche Anwendungskontexte hinweg und reduzieren den Bedarf an aufgabenspezifischer Modellierung. Der Übergang von spezialisierten Einzellösungen zu allgemein einsetzbaren Lernarchitekturen hat damit Reichweite und Einsatzmöglichkeiten KI-basierter Systeme erheblich erweitert.

Mit der Überführung dieser Modelle in reale Anwendungsszenarien treten Anforderungen in den Vordergrund, die über reine Leistungsfähigkeit hinausgehen. Gefordert wird ein Verhalten, das nachvollziehbar, stabil und an menschlichen Erwartungen orientiert ist, insbesondere in Situationen multimodaler Verarbeitung und Interaktion. Diese Entwicklung verschiebt den Forschungsschwerpunkt von der Skalierung von Modellkapazität hin zur Sicherstellung von Verlässlichkeit.

In dieser Dissertation wird Verlässlichkeit als eine Eigenschaft verstanden, die aus dem Zusammenwirken mehrerer Dimensionen entsteht: kontextuelle Verankerung, Stabilität der Repräsentationen, inhaltliche Treue sowie Steuerbarkeit. Diese Dimensionen entsprechen unterschiedlichen, jedoch miteinander gekoppelten Phasen des Modelllebenszyklus. Sie reichen von der multimodalen Fundierung sprachlicher Bedeutung über Anpassungsprozesse vortrainierter Repräsentationen bis hin zur Generierung von Ausgaben und deren gezielter Beeinflussung. Eine isolierte Betrachtung einzelner Phasen führt dabei zu charakteristischen Fehlermustern, etwa kontextunabhängiger Interpretation, Instabilität bei der Anpassung, Halluzinationen in generativen Prozessen oder sicherheitsbezogenen Eingriffen mit Bedeutungsverlust.

Vor diesem Hintergrund untersucht die Arbeit Verlässlichkeit als ein systemisches Entwurfsproblem. Entwickelt werden komplementäre Ansätze zur Stärkung kontextsensitiver multimodaler Verankerung bei gleichzeitiger Erhaltung semantischer Struktur, zur Regulierung generativer Prozesse auf Inferenzebene sowie zur Identifikation interpretierbarer Eingriffspunkte durch kognitiv inspirierte Analysen. Ziel ist nicht die isolierte Verbesserung einzelner Komponenten, sondern die koordinierte Behandlung unterschiedlicher Ursachen von Unzuverlässigkeit entlang der Daten-, Repräsentations- und Inferenzebenen.

Insgesamt zeigt sich, dass verlässliche Grundlagenmodelle nicht durch nachgelagerte Anpassungsmaßnahmen allein erreicht werden können, sondern als Eigenschaft des gesamten Modelllebenszyklus entworfen werden müssen. Verlässlichkeit entsteht durch die abgestimmte Verbindung von Verankerung, Anpassung, Inferenzverhalten und kontrollierbarer Steuerung und weist damit den Weg zu KI-Systemen, deren allgemeine Leistungsfähigkeit mit Transparenz, Vorhersagbarkeit und einer am Menschen orientierten Nutzbarkeit einhergeht.

Contents

List of Publications

Publications Forming the Basis of this Dissertation

Xintong Wang, Xiaoyu Li, Liang Ding, Sanyuan Zhao, and Chris Biemann. 2023. *Using Self-Supervised Dual Constraint Contrastive Learning for Cross-Modal Retrieval*. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023),* 2552–2559. Kraków, Poland: IOS Press.

Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024. *Probing Large Language Models from a Human Behavioral Perspective*. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING 2024,* 1–7. Torino, Italia: ELRA / ICCL.

Xintong Wang, Yixiao Liu, Jingheng Pan, Liang Ding, Longyue Wang, and Chris Biemann. 2025. *Chinese Toxic Language Mitigation via Sentiment Polarity Consistent Rewrites*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* 35695–35711. Suzhou, China: Association for Computational Linguistics.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. *Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding*. In *Findings of the Association for Computational Linguistics: ACL 2024,* 15840–15853. Bangkok, Thailand: Association for Computational Linguistics.

Xintong Wang, Jingheng Pan, Liang Ding, Longyue Wang, Longqin Jiang, Xingshan Li, and Chris Biemann. 2025. *CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models*. In *Findings of the Association for Computational Linguistics (ACL 2025),* 25507–25522. Vienna, Austria.

Xintong Wang*, Florian Schneider*, Özge Alaçam, Prateek Chaudhury, and Chris Biemann. 2022. *MOTIF: Contextualized Images for Complex Words to Improve Human Reading*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2468–2477. Marseille, France: European Language Resources Association.

Co-Authored Dissertation Related Publications

Abinew Ali Ayele, Nikolay Babakov, Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashraf Elnagar, Dayne Freitag, Maik Fröbe, Damir Korencic, Maximilian Mayerl, Daniil Moskovskiy, Animesh Mukherjee, Alexander Panchenko, Martin Potthast, Francisco Rangel, Naquee Rizwan, Paolo Rosso, Florian Schneider, Alisa Smirnova, Efstathios Stamatatos, Elisei Stakovskii, Benno Stein, Mariona Taulé, Dmitry Ustalov, **Xintong Wang**, Matti Wiegmann, Seid Muhie Yimam, and Eva Zangerle. 2024. *Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview*. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction,* 14959:231–259. Lecture Notes in Computer Science. Cham: Springer.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, **Xintong Wang**, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. *Multilingual and Explainable Text Detoxification with Parallel Corpora*. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025),* 7998–8025. Abu Dhabi, UAE: Association for Computational Linguistics.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, **Xintong Wang**, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. *Overview of the Multilingual Text Detoxification Task at PAN 2024*. In *Proceedings of the 2024 Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024),* 2432–2461. Grenoble, France.

Usman Naseem, Robert Geislinger, Juan Ren, Sarah Kohail, Rudy Alexandro Garrido Veliz, P Sam Sahil, Yiran Zhang, Idris Abdulmumin, Marco Antonio Stranisci, Özge Alaçam, Cengiz Acarturk, Aisha Jabr, Saba Anwar, Abinew Ali Ayele, Simona Frenda, Alessandra Teresa Cignarella, Elena Tutubalina, Oleg Rogov, Aung Kyaw Htet, **Xintong Wang**, Surendrabikram Thapa, Kritesh Rauniyar, Tanmoy Chakraborty, Md Arfeen Zeeshan, Dheeraj Kodati, Satya Keerthi, Sahar Moradizeyveh, Firoj Alam, Md Arid Hasan, Syed Ishtiaque Ahmed, Ye Kyaw Thu, Shantipriya Parida, Ihsan Ayyub Qazi, Lilian Diana Awuor Wanzare, Nelson Odhiambo Onyango, Clemencia Siro, Jane Wanjiru Kimani, Ibrahim Said Ahmad, Adem Chanie Ali, Martin Semmann, Chris Biemann, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2026. *POLAR: A Benchmark for Multilingual, Multicultural, and Multi-Event Online Polarization*. In *Findings of the Association for Computational Linguistics: ACL 2026,* 1–22. San Diego, California, USA: Association for Computational Linguistics.

Florian Schneider, Özge Alaçam, **Xintong Wang**, and Chris Biemann. 2021. *Towards Multi-Modal Text-Image Retrieval to Improve Human Reading*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (SRW at NAACL 2021),* 1–8. Online: Association for Computational Linguistics.

Anton Wiehe, Florian Schneider, Sebastian Blank, **Xintong Wang**, Hans-Peter Zorn, and Chris Biemann. 2022. *Language over Labels: Contrastive Language Supervision Exceeds Purely Label-Supervised Classification Performance on Chest X-Rays*. In *Proceedings of the 2022 Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop (SRW at AACL-IJCNLP 2022),* 76–83. Online: Association for Computational Linguistics.

List of Figures

List of Tables

List of Abbreviations

ACC . . . . . Accuracy
AI . . . . . Artificial Intelligence
BERT . . . . . Bidirectional Encoder Representations from Transformers
BLIP . . . . . Bootstrapping Language-Image Pre-training
BLEU . . . . . Bilingual Evaluation Understudy
CEFR . . . . . Common European Framework of Reference for Languages
CHAIR . . . . . Caption Hallucination Assessment with Image Relevance
CIDEr . . . . . Consensus-based Image Description Evaluation
CLIP . . . . . Contrastive Language-Image Pretraining
COMET . . . . . Crosslingual Optimized Metric for Evaluation of Translation
CWI . . . . . Complex Word Identification
DFG . . . . . German Research Foundation
FFN . . . . . Feed-Forward Network
FFD . . . . . First Fixation Duration
GD . . . . . Gaze Duration
GECO . . . . . Ghent Eye-Tracking Corpus
GLUE . . . . . General Language Understanding Evaluation
GPT . . . . . Generative Pre-trained Transformer
ICD . . . . . Instruction Contrastive Decoding
IR . . . . . Image Retrieval
ITI . . . . . Image-Text-Image Translation
L2 . . . . . Second Language
LLM . . . . . Large Language Model
LS . . . . . Lexical Simplification
LVLM . . . . . Large Vision-Language Model
MME . . . . . Multimodal Model Evaluation
MoE . . . . . Mixture of Experts
MOTIF . . . . . MultimOdal ConTextualized Images For Language Learners
MS COCO . . . . . Microsoft Common Objects in Context
MTurk . . . . . Amazon Mechanical Turk

MWE . . . . . Multi-Word Expression
NLP . . . . . Natural Language Processing
PEFT . . . . . Parameter-Efficient Fine-Tuning
POPE . . . . . Polling-based Object Probing Evaluation
POS . . . . . Part of Speech
QA . . . . . Question Answering
Q-Former . . . . . Querying Transformer
RLHF . . . . . Reinforcement Learning from Human Feedback
ROI . . . . . Region of Interest
RQ . . . . . Research Question
RTP . . . . . RealToxicityPrompts
SOTA . . . . . State of the Art
SUCCESSOR . . . . Unsupervised Dual Constraint Contrastive Learning for Cross-Modal Retrieval
TIT . . . . . Text-Image-Text Translation
TR . . . . . Text Retrieval
TRT . . . . . Total Reading Time
VCD . . . . . Visual Contrastive Decoding
ViT . . . . . Vision Transformer
VLM . . . . . Vision-Language Model
VLP . . . . . Vision-Language Pretraining
VQA . . . . . Visual Question Answering
WRA . . . . . Word-Region Alignment
ZuCo . . . . . Zurich Cognitive Language Processing Corpus

Statement on Software and AI Assistance

This dissertation was prepared and typeset in LaTeX using Overleaf, based on the Universitat Hamburg LT Thesis Template. Figures were created in Microsoft PowerPoint.

During the writing and revision process, Grammarly (web version) was used for grammar and spelling suggestions. The ChatGPT desktop application was used only to generate suggestions for grammar, spelling, or style corrections and to provide high-level editorial feedback on clarity and structure; any accepted changes were incorporated manually by the author. It was not used to generate dissertation content or running text.

Google Gemini 3.1 Pro was used to create an initial draft of the German translation of the abstract. This draft was then manually corrected and polished by the author.

All scientific arguments, literature selection, factual claims, methodological decisions, data analyses, and final formulations presented in this dissertation were made by the author.

Funding Acknowledgments

This research was supported by the German Research Foundation (DFG) through the Collaborative Research Centre SFB 169 Crossmodal Learning: Adaptivity, Prediction and Interaction, as well as by the Excellence Funds of the University of Hamburg. These funding sources provided the institutional and infrastructural support that enabled the research presented in this dissertation.

All models are wrong, but some are useful.

— George E. P. Box (1979)

1 Introduction

Contents

1.1 Background and Motivation

Across many domains, artificial intelligence is now embedded in everyday computing, supporting tasks from search and translation to writing, design, and decision support. As these systems enter increasingly diverse settings, the expectations placed on them extend beyond whether they can perform a given task to how reliably and meaningfully they do so.

Artificial intelligence has undergone a substantial shift from systems designed for narrowly specified tasks to pretrained models that can support a broad range of downstream applications. Earlier paradigms typically optimized separate models for individual problems such as classification, translation, or question answering, whereas large-scale pretraining has made it possible to learn representations that transfer across tasks with comparatively little additional supervision (Bommasani et al., 2021; Brown et al., 2020). This transition has given rise to foundation models, whose broad adaptability has reshaped how language processing, knowledge access, and model deployment are approached in contemporary AI research and practice (Bommasani et al., 2021; Achiam et al., 2023). Rather than treating each task as an isolated modeling problem, current systems increasingly begin from general pretrained architectures that are subsequently adapted, instructed, or specialized for particular downstream settings.

A particularly important extension of this transition is the rise of multimodal foundation models that jointly process language and visual information. Large-scale vision–language pretraining has shown that visual and textual representations can be aligned in shared or tightly coupled spaces, enabling models to transfer across retrieval, classification, and captioning tasks (Radford et al., 2021; J Li et al., 2022). More recent large vision–language models (LVLMs) (Haotian Liu et al., 2023; J Li et al., 2023; W Dai et al., 2023; D Zhu et al., 2024) build on this foundation by combining visual perception with generative language modeling, thereby extending the scope of foundation models from text-only interaction to multimodal reasoning, instruction following, and response generation. This development is especially consequential because many real-world settings in which AI systems are used, including tasks such as visually grounded question answering and document interpretation as well as broader multimodal assistant settings, require models to integrate perceptual evidence with linguistic context rather than operate over text alone.

Yet increased capability does not by itself guarantee trustworthy behavior. When foundation models are deployed in real-world settings, they frequently exhibit recurrent failure modes, including generating content that is unsupported by the input, producing harmful or inappropriate outputs, and responding in ways that remain difficult to interpret or control (Bender et al., 2021; Weidinger et al., 2021; Liang et al., 2023; L Huang et al., 2023; Hanchao Liu et al., 2024; Song et al., 2026). These failures are not merely isolated defects of individual models, but symptoms of deeper structural tensions between model performance, reliability, and dependable deployment. The resulting gap between capability-oriented evaluation and trustworthy behavior has therefore become a central problem in the broader transition from experimental success to practical adoption.

In broader discussions of AI, trust is often treated as a relational notion involving users, institutions, and deployment contexts (Jacovi et al., 2021), whereas trustworthiness is more naturally used to refer to properties of systems that make such trust warranted or unwarranted (Miedema et al., 2026). The present dissertation adopts a deliberately scoped view of trustworthiness at the level of model behavior in multimodal generative foundation models. The aim is not to provide a comprehensive account of trustworthy AI across fairness, privacy, security, governance, and other broader socio-technical concerns (Díaz-Rodríguez et al., 2023; Song et al., 2026), but to examine whether model behavior remains grounded in its inputs, stable under adaptation, faithful to conditioning information, and selectively controllable at inference time. Under this scoped interpretation, trustworthiness is treated as a functional and behavioral property of model operation, one that can be analyzed through recurrent failure modes and targeted interventions rather than only through abstract principles (Song et al., 2026).

To analyze these issues systematically, the dissertation adopts a *trustworthiness pipeline* as its central organizing perspective. The term is used here in a conceptual rather than purely systems-engineering sense: it refers to the connected stages at which unreliability may arise and at which interventions may be applied in foundation model development and use (Miedema et al., 2026; Song et al., 2026). In the present work, these stages span data construction and supervision, representation learning and multimodal alignment, adaptation to new tasks or domains, inference-time generation, and behavior regulation or control. Under this view, trustworthiness is not treated as a single attribute attached only to final outputs, but as a property that is supported or undermined across connected stages of model development and deployment.

Within this pipeline, unreliability often appears not as an isolated defect of a single component, but as a recurring disconnection between stages that should remain coordinated. In the context of this dissertation, three broad forms of disconnection are especially consequential: a gap between perceptual evidence and contextual interpretation, instability introduced when pretrained representations are adapted to new tasks or domains, and a breakdown between internal model behavior and the conditions that should govern generation and control at inference time. These disconnections give rise to familiar failure phenomena, including multimodal hallucination, semantic drift under adaptation, and safety interventions that may reduce harmful expression at the cost of weakening meaning or communicative intent (Rohrbach et al., 2018; Y Li et al., 2023; Bang et al., 2025; Gehman et al., 2020; H Zhao et al., 2024; Song et al., 2026). The dissertation therefore approaches trustworthiness not as a single undifferentiated property, but as a structured problem space whose recurring tensions can be operationalized more precisely in the analytical framework developed below. In particular, the first two disconnections motivate the dimensions of groundedness and alignment stability. By contrast, breakdowns at inference time appear in two distinct forms: failures of faithfulness to conditioning information and failures of controllability and safety in behavior regulation.

The title of this dissertation foregrounds three complementary lenses on trustworthy foundation models. *Vision* refers to the role of perceptual evidence in grounding multimodal understanding and constraining faithful generation. *Language* refers to the semantic structures that models must preserve across adaptation, generation, and controlled language behavior. *Gaze* refers to human cognitive signals that provide an interpretable basis for identifying where model behavior can be analyzed and selectively steered. Taken together, vision, language, and gaze define the central perspective of the thesis: trustworthiness depends on keeping perceptual grounding, linguistic structure, and human-informed control tightly connected.

1.2 Operationalizing Trustworthiness

Building on the scoped and pipeline-oriented account of trustworthiness introduced above, this section operationalizes the framework into four analytical dimensions. This decomposition makes recurrent forms of unreliability more precisely identifiable across different stages of foundation model development and use, while preserving the connections among them. It thus provides a common basis for analyzing failure modes and locating intervention points within a coherent trustworthiness perspective.

These dimensions are derived from the recurring disconnections identified in the preceding section and are ordered broadly along the trustworthiness pipeline. The first two disconnections map directly onto groundedness and alignment stability, since they concern, respectively, the relation between perceptual evidence and contextual interpretation and the preservation of semantic structure under adaptation. The third disconnection, however, separates into two distinct inference-time questions: whether outputs remain supported by their conditioning information, and whether model behavior can be regulated in a selective and meaning-preserving way. For this reason, the dissertation distinguishes between faithfulness and controllability and safety as closely related but analytically distinct trustworthiness requirements.

Groundedness: Groundedness refers to the extent to which model representations and interpretations remain anchored to the input signals and their contextual environment (Harnad, 1990). In multimodal settings, this requires establishing reliable relations between linguistic expressions and the perceptual and contextual evidence that supports their interpretation, rather than relying only on coarse global correspondence or statistical association (Plummer et al., 2015; Hudson and Manning, 2019). Groundedness becomes a trustworthiness issue when a model produces semantically plausible yet contextually unsupported interpretations, especially in cases where the same expression admits different meanings under different usage conditions.

In this dissertation, groundedness is treated as a property that must be supported both by data construction and by model interaction with multimodal inputs. Operationally, it is examined through context-sensitive multimodal resource design and through human judgments of whether visual evidence genuinely supports the interpretation of a target expression in context (Schneider et al., 2021; **Wang* et al.**, 2022). On this view, trustworthy multimodal understanding depends not only on representational capacity, but also on whether models receive supervision that makes contextual support explicit and reliably interpretable.

Alignment Stability: Alignment stability concerns the preservation of semantic structure when foundation models are adapted to new domains, tasks, or modalities, especially the cross-modal and intra-modal relationships established during pretraining. Although large-scale pretraining yields broadly transferable representations, subsequent fine-tuning or efficient adaptation may distort the relationships encoded in those representations, weakening cross-modal correspondence or disrupting intra-modal coherence (H Li et al., 2024). In this dissertation, alignment stability therefore refers not simply to successful adaptation, but to whether specialization can be achieved without eroding the semantic structure on which reliable transfer depends (Houlsby et al., 2019; Hu et al., 2022; Shihab et al., 2026; K Yao et al., 2026).

This becomes a trustworthiness issue when adaptation produces task-specific competence at the cost of representational integrity. A model may appear effective within a narrow downstream setting while becoming less consistent, less transferable, or less predictable outside that setting. From this perspective, stable alignment is what allows foundation models to extend their competence without sacrificing the reliability of their learned semantic relationships. Operationally, the dissertation studies this dimension in efficient adaptation settings where downstream specialization must be achieved while preserving both cross-modal correspondence and intra-modal consistency.

Faithfulness: Faithfulness describes the degree to which model outputs remain supported by the conditioning information available at inference time. In multimodal generation, this includes maintaining consistency with visual evidence, textual instructions, and other contextual signals that should govern what may be validly produced. Whereas groundedness concerns whether multimodal interpretation is anchored to appropriate evidence, faithfulness concerns whether generated outputs remain supported by those conditioning signals during decoding. It becomes a trustworthiness issue when a model produces fluent and plausible content that is not in fact supported by the inputs on which the response is supposed to depend.

A prominent manifestation of unfaithfulness is hallucination, in which models generate unsupported or fabricated content (Rohrbach et al., 2018; Y Li et al., 2023; Hanchao Liu et al., 2024; Bang et al., 2025). However, faithfulness in this dissertation is treated more broadly as an inference-time property of trustworthy generation. It is

shaped not only by model parameters, but also by the mechanisms through which learned priors, contextual conditioning, and external evidence are balanced during decoding (XL Li et al., 2023; Leng et al., 2024; Su et al., 2025; C Li et al., 2026). Operationally, this dimension is studied through hallucination-oriented evaluation of multimodal outputs and through inference-time controls that suppress unsupported content without requiring retraining of the base model.

Controllability and Safety: Controllability and safety refer to the capacity to guide model behavior in a transparent, selective, and task-sensitive manner while preserving the communicative intent encoded in the input. In this dissertation, safety is treated not as a purely suppressive objective, but as a constrained form of controllability: models should be steerable toward acceptable behavior without unnecessary distortion of meaning, tone, or pragmatic function. This becomes a trustworthiness issue when interventions achieve surface-level acceptability only by undermining the semantic or pragmatic properties that the original expression was meant to convey.

Traditional safety interventions often rely on blocking, neutralization, or coarse rewriting strategies that reduce harmful expression while also weakening sentiment, expressiveness, or intended meaning (Gehman et al., 2020; Logacheva et al., 2022; **Wang**, Liu, et al., 2025). The dissertation instead operationalizes this dimension through tasks that jointly assess toxicity reduction and preservation of sentiment or meaning, as well as through selective interventions that target specific internal components rather than globally altering outputs. On this view, trustworthy control requires interpretable mechanisms for modulating undesirable behavior without sacrificing meaning, tone, or pragmatic function.

Together, these four dimensions—**groundedness, alignment stability, faithfulness, and controllability and safety**—form the analytical backbone of the dissertation's account of trustworthy foundation models. Each dimension corresponds to a distinct locus of unreliability across the trustworthiness pipeline, yet they are tightly coupled in practice: insufficient grounding can place pressure on subsequent adaptation, unstable representations can undermine faithfulness during generation, and coarse forms of control can turn safety interventions into meaning-distorting transformations. These interdependencies motivate the research challenges developed in the next section.

1.3 Research Challenges

The operational dimensions introduced above identify where trustworthiness breaks down across the foundation model pipeline; the present section reformulates these breakdowns as concrete research challenges. Rather than treating groundedness, alignment stability, faithfulness, and controllability and safety as abstract desiderata, the dissertation approaches them as recurring practical tensions that arise during data construction, adaptation, generation, and behavioral regulation. In this way, the challenges below define the immediate problem space from which the dissertation's research questions follow.

Challenge I: Establishing and Evaluating Context-Sensitive Grounding. Foundation models are often trained on large-scale paired data in which relations between language and visual information are learned through broad statistical regularities. While this paradigm is effective for capturing general correspondence, it does not by itself ensure that models learn which perceptual evidence is actually relevant to a particular

expression under its immediate contextual conditions. As a result, a model may rely on visually plausible information that is globally related to the text while remaining locally misleading for the interpretation that the context requires.

This challenge becomes especially important in human-centered settings such as reading support, language learning, and multimodal disambiguation, where the value of perceptual evidence depends on whether it genuinely supports the intended interpretation of a target expression. The problem is therefore not only to acquire multimodal data, but to construct and evaluate resources that explicitly encode contextual relevance rather than generic visual relatedness. Without such grounding, downstream interpretation and generation inherit ambiguities that cannot be resolved reliably at later stages of the pipeline.

Challenge II: Preserving Semantic Structure Under Efficient Adaptation. Adapting foundation models to new tasks or domains is essential for practical deployment, yet adaptation must not come at the cost of the semantic structure that underlies their broader transferability. Full-parameter fine-tuning can overwrite or distort pretrained knowledge, while lightweight adaptation techniques may improve efficiency without adequately constraining how representations shift across modalities or tasks. As a result, models may achieve strong performance within a narrow adaptation setting while becoming less coherent, less transferable, or less reliable outside it.

The central challenge is therefore to enable efficient specialization without semantic drift. Trustworthy adaptation requires mechanisms that preserve the structural relationships established during pretraining while still allowing targeted refinement for downstream demands. In the multimodal case, this includes maintaining both cross-modal correspondence and intra-modal consistency, so that efficiency gains do not come at the cost of representational integrity or dependable behavior.

Challenge III: Maintaining Faithful Generation Under Competing Signals. Even when multimodal representations are well aligned, generation-time behavior can still diverge from the conditioning information available at inference time. During decoding, foundation models must reconcile multiple influences, including contextual conditioning, learned priors, and statistical regularities acquired during pretraining. When these influences are imbalanced, models may produce fluent yet unsupported statements, of which hallucination is a prominent example.

The challenge is therefore not generation in itself, but faithful generation under competing signals. Trustworthy multimodal systems require inference-time mechanisms that regulate how external evidence and internal knowledge are balanced during output production. Without such mechanisms, parameter learning alone does not guarantee that generated content will remain supported by the conditioning information on which it is supposed to depend.

Challenge IV: Achieving Safe, Meaning-Preserving, and Interpretable Control. Safety interventions are often implemented through blocking, rewriting, or other suppression-based strategies that reduce overtly harmful expression at the surface level. While such approaches may lower immediate risk, they can also distort meaning, sentiment, tone, or pragmatic intent, producing responses that are formally acceptable yet communicatively misaligned. The challenge is therefore not only to make model behavior safer, but to do so without sacrificing the semantic and expressive properties that the input was meant to convey.

At the same time, trustworthy control must be interpretable and targeted rather than opaque and globally applied. If undesirable behavior can be changed only through

coarse interventions, it becomes difficult to determine what aspect of behavior has been modified, why it has changed, and whether the modification remains faithful to the original communicative goal. The challenge, then, is to develop mechanisms that support selective and meaning-preserving regulation of model behavior while exposing intervention points that can be identified and steered precisely.

Taken together, these four challenges define the immediate research problem space of the dissertation. They arise at different points along the trustworthiness pipeline, yet they cannot be resolved independently: weak grounding propagates ambiguity, unstable adaptation weakens reliable downstream behavior, and coarse forms of control can distort behavior rather than regulate it precisely. The following research questions make these tensions more explicit in a form that can be investigated empirically.

1.4 Research Questions

The challenges outlined above motivate a set of research questions that guide the dissertation. Each question focuses on a specific trustworthiness problem within the foundation model pipeline and frames it in a form that can be investigated through the empirical studies that follow.

The first four questions correspond to the four challenge areas developed in the preceding section: context-sensitive grounding, stability under adaptation, faithful generation, and safe and meaning-preserving control. The fifth asks how interpretable and targeted intervention can be identified and applied across these challenge areas. In this way, the final question treats cognitively informed steering not as a separate trustworthiness dimension, but as a cross-cutting problem of locating and operationalizing control within the broader pipeline.

Figure 1.1 provides a compact visual map of how these research questions and chapter-level studies are positioned across the trustworthiness pipeline.

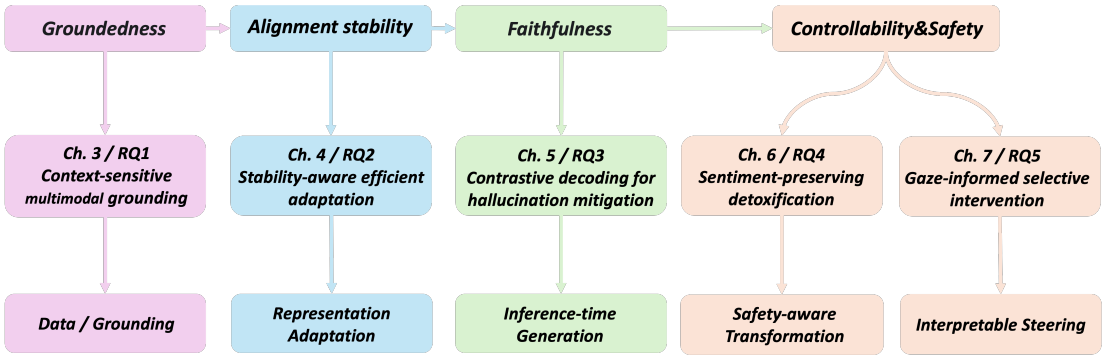


Figure 1.1: Overview of the dissertation across the trustworthiness pipeline. The figure maps the four operational dimensions introduced in Chapter 1 to the chapter-level studies in Chapters 3–7 and their primary intervention points. The split under controllability and safety indicates two complementary intervention paths, emphasizing that the dissertation contributes coordinated evidence across the pipeline rather than a single end-to-end integrated system.

RQ1: Can context-sensitive multimodal resource construction improve grounded understanding in human-centered settings? Reliable multimodal behavior depends not only on model architecture but also on whether supervision makes contextual relevance explicit during training and evaluation. If visual information is only weakly

tied to linguistic context, models may learn superficial associations that fail to resolve ambiguity or support human-centered comprehension. This question therefore asks whether context-sensitive multimodal resource construction can strengthen grounded semantic relationships by jointly modeling textual difficulty, perceptual support, and contextual usage.

RQ2: Can foundation models be efficiently adapted while preserving the stability of their pretrained semantic structure? Practical deployment requires adapting pretrained models to domain-specific tasks, yet such adaptation must avoid degrading the representational integrity that underlies their broader generalization ability. This question asks whether efficient adaptation mechanisms can extend model capabilities while preserving the cross-modal alignment and intra-modal coherence established during large-scale pretraining.

RQ3: Can hallucinations in multimodal generation be mitigated through inference-time mechanisms rather than additional training? Hallucinations reveal a mismatch between learned priors and the conditioning information available during generation, often emerging when decoding dynamics privilege internal knowledge over external evidence. This question therefore asks whether faithfulness can be improved by directly regulating decoding dynamics, rather than relying on further parameter updates or retraining.

RQ4: Can safety interventions remove harmful content while preserving communicative intent and affective meaning? Many safety interventions reduce harmfulness by suppressing, neutralizing, or rewriting content in ways that also weaken sentiment, tone, or intended meaning. This question asks whether trustworthy control can distinguish harmful expression from legitimate affect, so that safety is achieved without collapsing diverse communicative intent into affectively flattened or semantically degraded outputs.

RQ5: Can human cognitive signals provide interpretable guidance for identifying and selectively steering internal model behavior? Targeted control depends not only on whether model behavior can be changed, but also on whether meaningful intervention points can be identified in the first place. Human cognitive signals such as gaze provide interpretable evidence about relevance and attentional priority, and may therefore help localize where selective control should be applied. This question asks whether such signals can help identify and implement precise, selective interventions for steering foundation model behavior.

Taken together, these five questions define a progression across the trustworthiness pipeline, from grounded understanding and stable adaptation to faithful generation and safe and meaning-preserving control. The first four questions address distinct loci at which trustworthiness may break down, whereas the fifth asks how interpretable and targeted intervention can be identified and operationalized across them. Together, they define the dissertation's empirical research program for establishing, preserving, and improving trustworthy behavior in foundation models.

1.5 Contributions

Guided by the research questions above, the dissertation makes five contributions that collectively advance a pipeline-oriented account of trustworthy foundation models. Rather than proposing a single end-to-end technique, the work develops complementary mechanisms spanning context-sensitive resource construction, stable adaptation,

inference-time faithfulness regulation, safe and meaning-preserving control, and cognitively informed targeted intervention. The first four contributions correspond to the four trustworthiness loci defined above, while the fifth addresses interpretable intervention as a cross-cutting control problem. Together, they are substantiated through the studies presented in Chapters 3–7.

Table 1.1 summarizes how these contributions map to the research questions and dissertation chapters.

Table 1.1: Mapping of research questions to dissertation chapters and contributions.

RQ	Trustworthiness focus	Chapter	Publication basis	Main contribution
RQ1	Groundedness	Chapter 3	MOTIF (LREC 2022)	Context-sensitive multimodal resources for visually supported reading.
RQ2	Alignment stability	Chapter 4	Dual constraint contrastive learning (ECAI 2023)	Stable cross-modal adaptation via inter-modal and intra-modal constraints.
RQ3	Faithfulness	Chapter 5	Instruction contrastive decoding (ACL Findings 2024)	Inference-time mitigation of LVLM hallucinations without additional training.
RQ4	Safe and meaning-preserving control	Chapter 6	ToxiRewriteCN (EMNLP 2025)	Sentiment-preserving detoxification for safe, meaning-preserving generation.
RQ5	Interpretable and targeted control	Chapter 7	Probing LLMs and CogSteer (NeusymBridge@LREC-COLING 2024; ACL Findings 2025)	Gaze-informed layer selection and selective intervention for efficient semantic steering.

C1: A Context-Sensitive Resource Construction Framework for Grounded Multimodal Understanding. To address the challenge of acquiring reliable multimodal supervision (RQ1), this work introduces a resource construction methodology that explicitly models the relationship between linguistic complexity, contextual usage, and visual depictability. By selecting focus expressions based on their semantic difficulty and grounding requirements, and by pairing them with contextually appropriate visual evidence, the resulting resource enables the study of how perceptual support can facilitate accurate interpretation rather than merely provide illustrative correlation. This contribution establishes a principled approach to grounded multimodal resource design and demonstrates how context-aware alignment can support the analysis of grounding behavior in human-centered scenarios (**Wang* et al.**, 2022).

C2: An Efficient Dual-Constraint Adaptation Strategy for Alignment Stability. Addressing RQ2, the dissertation proposes a lightweight adaptation paradigm that preserves pretrained semantic structure while enabling domain specialization. The method introduces a dual-constraint objective that simultaneously enforces inter-modal correspondence and intra-modal consistency, allowing models to refine their representations without disrupting the global geometry acquired during pretraining. This approach shows that efficient adaptation can preserve key aspects of pretrained semantic structure

while enabling domain specialization, thereby reconciling computational practicality with representational stability (**Wang et al.**, 2023).

C3: An Inference-Time Contrastive Decoding Mechanism for Improving Faithfulness. To investigate RQ3, this work develops an inference-only decoding strategy designed to reduce hallucinations in multimodal generation. Instead of retraining model parameters, the method contrasts alternative decoding distributions induced by controlled instruction perturbations, enabling the model to suppress unsupported concepts while retaining plausible outputs. This contribution demonstrates that faithfulness can be enhanced through principled manipulation of inference dynamics, highlighting the critical role of generation-time control in trustworthy modeling (**Wang**, Pan, Ding, and Biemann, 2024).

C4: A Sentiment-Preserving Detoxification Paradigm for Safe and Meaning-Preserving Control. In response to RQ4, the dissertation introduces a framework for detoxification that explicitly distinguishes harmful expression from affective intent. Through the construction of a human-verified rewriting benchmark and a multi-stage annotation protocol, the work shows that safety interventions can be formulated as constrained semantic transformations rather than neutralizing edits. Empirical analyses reveal systematic trade-offs between toxicity mitigation and emotional fidelity, providing new insights into how controllable generation must account for pragmatic meaning (**Wang**, Liu, et al., 2025).

C5: A Cognition-Inspired Selective Layer Intervention Framework for Interpretable and Targeted Control. Finally, addressing RQ5 as a cross-cutting question of interpretable and targeted intervention, the dissertation presents a cognition-inspired steering framework that leverages human reading signals to identify functionally relevant layers within large language models. By correlating model representations with gaze-derived indicators and performing targeted interventions at selected layers, the method enables precise, efficient steering with minimal parameter modification. This contribution demonstrates how cognitively grounded analysis can help localize selectively steerable components within neural architectures, linking interpretability with practical controllability (**Wang**, Pan, Ding, et al., 2025).

Taken together, these contributions show that trustworthiness in foundation models is not the result of a single architectural innovation, but of coordinated design across data construction, representation alignment, generation, and behavior regulation. The dissertation therefore advances a unified perspective in which grounded supervision, stable adaptation, faithful generation, and safe and meaning-preserving control define the principal trustworthiness loci, while cognitively informed targeted intervention provides a cross-cutting mechanism for analyzing and steering them.

1.6 Contributions to the Underlying Papers

This dissertation is based on six first-authored publications, including one co-first-authored paper. The summaries below clarify my leading role in the conception, design, implementation, analysis, and writing of the underlying papers, while briefly noting the principal contributions of co-authors where relevant.

In the paper *MOTIF: Contextualized Images for Complex Words to Improve Human Reading* (**Wang* et al.**, 2022), I led the formulation of the research problem, the dataset design, the annotation protocol, the analysis, and the drafting of the paper. Özge Alaçam

contributed to the problem formulation and paper revision. Florian Schneider conducted the crowdsourcing experiments and the context-dependent image retrieval component, and contributed to the writing of the corresponding sections. Chris Biemann provided feedback that helped improve the work.

In the paper *Using Self-Supervised Dual Constraint Contrastive Learning for Cross-Modal Retrieval* (**Wang et al.**, 2023), I led the problem formulation, method design, implementation, experiments, and paper writing. Xiaoyu Li conducted the error analysis. Liang Ding contributed proofreading and revision feedback. Sanyuan Zhao contributed to discussion of the idea. Chris Biemann supervised the work and provided proofreading and feedback on the paper.

In the paper *Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding* (**Wang**, Pan, Ding, and Biemann, 2024), I led the problem formulation, method design, implementation, experiments, and paper writing. Jingheng Pan contributed the instruction sensitivity analysis with positive and negative prompts. Liang Ding participated in discussions of the core idea. Chris Biemann provided proofreading and extensive feedback during revision.

In the paper *Chinese Toxic Language Mitigation via Sentiment Polarity Consistent Rewrites* (**Wang**, Liu, et al., 2025), I led the problem formulation, dataset design, annotation protocol, analysis, and paper writing. Yixiao Liu contributed to discussion and data filtering. Jingheng Pan contributed to discussions and to revisions of the presentation of the tables. Liang Ding and Longyue Wang participated in discussion of the work. Chris Biemann provided proofreading and feedback on the paper.

In the paper *Probing Large Language Models from a Human Behavioral Perspective* (**Wang**, Li, et al., 2024), I led the problem formulation, method design, implementation, experiments, and paper writing. Xiaoyu Li contributed to the preparation of the eye-tracking data. Xingshan Li contributed to the discussion of the work. Chris Biemann supervised the project and provided proofreading and feedback on the paper.

In the paper *CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models* (**Wang**, Pan, Ding, et al., 2025), I led the problem formulation, method design, implementation, experiments, and paper writing. Jingheng Pan contributed to the GLUE benchmark evaluation. Xingshan Li provided cognitive insights during the discussion stage. Liang Ding and Longyue Wang contributed to discussion of the work. Chris Biemann supervised the project and provided proofreading and feedback on the paper.

1.7 Structure of the Dissertation

The dissertation is organized to address the research questions outlined above through a sequence of studies that examine trustworthiness across different stages of the foundation model pipeline.

Chapter 2 reviews related work on large language models, vision–language modeling, multimodal alignment, controllable generation, and cognitively inspired approaches to model interpretability and steering. It situates the present work within ongoing efforts to improve the reliability and transparency of foundation models and highlights the limitations that motivate the proposed framework.

Chapters 3–7 present the core contributions of this dissertation. Chapters 3–6 each build on a previously published study, while Chapter 7 consolidates and extends two

related first-authored studies. Chapter 3 investigates context-sensitive multimodal grounding through a dataset for visually informed language understanding. Chapter 4 introduces a dual-constraint adaptation method for alignment stability in vision–language representations. Chapter 5 addresses hallucination in multimodal generation with an inference-time contrastive decoding strategy. Chapter 6 explores sentiment-preserving detoxification as a meaning-preserving approach to safe text generation. Chapter 7 presents a cognition-inspired selective layer intervention framework for interpretable and targeted steering of large language models. Chapter 8 concludes the dissertation by summarizing the findings, discussing their limitations, and outlining future directions for trustworthy foundation models, especially the integration of human cognitive signals, multimodal grounding, and controllable inference mechanisms in unified learning systems.

You shall know a word by the company it keeps.

— J. R. Firth (1957)

2

Technical and Research Background

Contents

This chapter provides the technical and research background for the dissertation. It first establishes the core architectural and methodological foundations on which the subsequent chapters build. These foundations include transformer-based modeling, multimodal alignment and fusion, adaptation regimes, inference-time generation, decoding, and control, and selected human cognitive signals, particularly gaze during reading. The chapter then selectively reviews the lines of work that most directly support the studies presented in Chapters 3–7. Rather than offering a comprehensive survey of all topics related to trustworthy AI, it focuses on the research areas that define the scope of this thesis: multimodal grounding, efficient adaptation, faithful generation, safety-aware control, and interpretable intervention. This framing reflects the dissertation's aim of developing a connected perspective on trustworthiness across the foundation model pipeline, while also preparing the ground for later discussions of how human cognitive signals can inform interpretable model steering.

2.1 Technical and Methodological Foundations

2.1.1 Transformer-Based Foundation Models

The language and multimodal foundation models most relevant to this dissertation are based predominantly on the transformer architecture, which models sequences through contextualized vector representations rather than through recurrence (Vaswani et al., 2017; Bommasani et al., 2021). Let an input sequence be denoted by $x = (x_1, \dots, x_n)$, where x_t is the token at position $t \in \{1, \dots, n\}$. Here, a token may correspond to a word, a subword unit, or punctuation, depending on the tokenizer used by the model. Each token is first mapped to a learned embedding vector $e_t \in \mathbb{R}^d$, where d denotes the model dimension. Because attention by itself is content-based and does not encode sequence order, each token embedding is combined with positional information before entering the first transformer layer. Denoting the positional encoding at position t by $p_t \in \mathbb{R}^d$, the initial hidden representation matrix is

$$H^{(0)} = \begin{bmatrix} e_1 + p_1 \\ \vdots \\ e_n + p_n \end{bmatrix} \in \mathbb{R}^{n \times d}, \tag{2.1}$$

where the t-th row of $H^{(0)}$ represents the input vector at position t. Positional information is necessary because word order changes meaning: for example, “dog bites man” and “man bites dog” contain the same lexical items, but they encode different semantic roles because the positions of the two nouns are reversed.

In the original transformer, positional information is introduced through deterministic sinusoidal functions (Vaswani et al., 2017). Let $p_{t,j}$ denote the j-th component of the positional encoding at position t, and let $j = 0, \dots, \frac{d}{2} - 1$. Using the present subsection’s 1-based token indexing, the sinusoidal encoding is

$$p_{t,2j} = \sin\left(\frac{t-1}{10000^{2j/d}}\right), \qquad p_{t,2j+1} = \cos\left(\frac{t-1}{10000^{2j/d}}\right). \tag{2.2}$$

These functions assign each position a distinct vector while preserving smooth relationships across nearby and distant positions. Many later transformer variants replace this scheme with learned absolute position embeddings, as in BERT (Devlin et al., 2019), or with relative positional representations that encode pairwise token offsets directly in the attention mechanism (Shaw et al., 2018). The underlying purpose remains the same: to inject order information that pure attention would otherwise miss.

Let $H^{(\ell-1)} \in \mathbb{R}^{n \times d}$ denote the hidden representations entering layer ℓ. In a self-attention sublayer, these representations are projected into query, key, and value matrices, denoted by Q, K, and V:

$$Q = H^{(\ell-1)} W_Q, \qquad K = H^{(\ell-1)} W_K, \qquad V = H^{(\ell-1)} W_V, \tag{2.3}$$

where $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, and $W_V \in \mathbb{R}^{d \times d_v}$ are learned projection matrices, d_k is the query/key dimension, and d_v is the value dimension. Scaled dot-product attention then computes

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \tag{2.4}$$

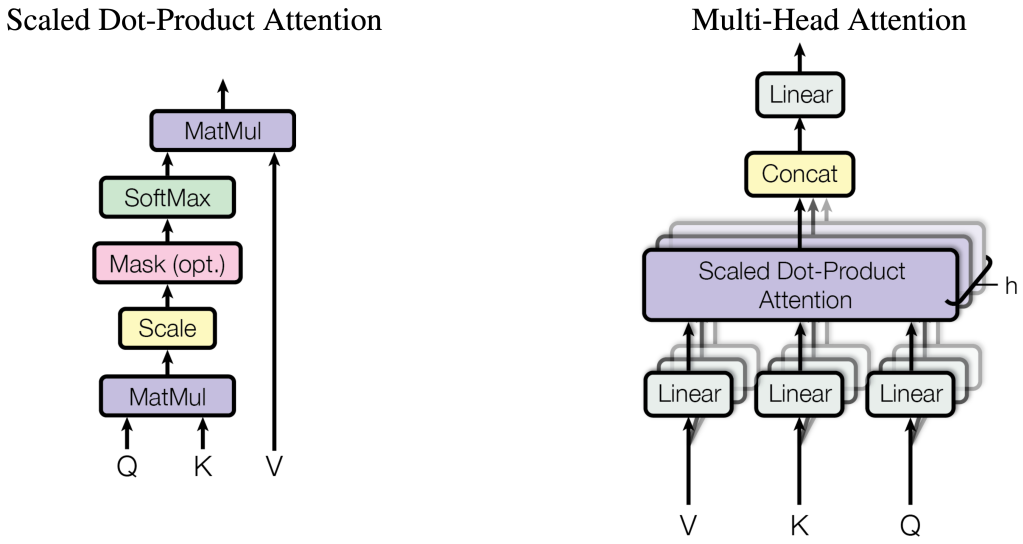


Figure 2.1: Illustration of scaled dot-product attention and multi-head attention. The left panel shows how query-key similarity scores are scaled, normalized, and used to weight value vectors, while the right panel shows how multiple attention heads are computed in parallel and recombined through concatenation and linear projection. Adapted from (Vaswani et al., 2017).

The matrix $QK^\top \in \mathbb{R}^{n\times n}$ contains pairwise similarity scores between token positions. The softmax function is applied row-wise, so each row becomes a distribution over the positions from which the current token should gather information. Dividing by $\sqrt{d_k}$ prevents the dot products from becoming excessively large in high dimensions, which would otherwise make the softmax distribution overly peaked and the gradients correspondingly small. Multiplication by V then yields a weighted combination of value vectors, producing updated token representations that depend on the broader context rather than on isolated word identities alone.

This contextualization is one of the defining strengths of transformers. A token such as "bank" should not receive the same representation in "the bank approved the loan" and "the canoe reached the bank," because the surrounding context indicates different senses. Self-attention supports this behavior by allowing the representation at each position to be recomputed in light of the other tokens in the sequence.

In practice, transformers use *multi-head attention*, in which several attention heads are computed in parallel. For head $i \in \{1, \dots, h\}$,

$$\text{head}_i \;=\; \text{Attention}\big(H^{(\ell-1)} W_i^Q,\; H^{(\ell-1)} W_i^K,\; H^{(\ell-1)} W_i^V\big)\,, \tag{2.5}$$

$$\text{MultiHead}\big(H^{(\ell-1)}\big) \;=\; \text{Concat}(\text{head}_1, \dots, \text{head}_h)\, W_O, \tag{2.6}$$

where h is the number of attention heads and W_O is an output projection matrix. Multi-head attention is not simply repeated attention; rather, it allows different subspaces of the representation to focus on different kinds of dependencies. For example, one head may emphasize local modifiers, another long-distance dependencies, and another entity-level associations. Figure 2.1 illustrates both scaled dot-product attention and its multi-head extension.

A transformer layer consists of more than attention alone. After multi-head attention, the model applies a residual connection and layer normalization, followed by a position-wise feed-forward network (FFN), a second residual connection, and a second

normalization step (Vaswani et al., 2017; Ba et al., 2016). A standard encoder-style update can be written as

$$\widetilde{H}^{(\ell)} = \mathrm{LayerNorm}\big(H^{(\ell-1)} + \mathrm{MultiHead}\big(H^{(\ell-1)}\big)\big), \tag{2.7}$$

$$H^{(\ell)} = \mathrm{LayerNorm}\big(\widetilde{H}^{(\ell)} + \mathrm{FFN}\big(\widetilde{H}^{(\ell)}\big)\big), \tag{2.8}$$

where LayerNorm denotes layer normalization and FFN is the same feed-forward transformation applied independently to each token representation. A standard position-wise FFN can be written for a single token representation $h \in \mathbb{R}^d$ as

$$\mathrm{FFN}(h) = \phi(hW_1 + b_1)W_2 + b_2, \tag{2.9}$$

where W_1 and W_2 are learned weight matrices, b_1 and b_2 are bias terms, and ϕ is an element-wise nonlinear activation function. Residual connections help preserve and refine information across depth, while normalization stabilizes optimization. The FFN introduces additional nonlinear transformation after contextual information has been aggregated through attention.

Transformer architectures are commonly instantiated in encoder-style, decoder-style, or encoder-decoder forms. In encoder-style models such as BERT, attention is typically bidirectional, allowing each token to attend to positions on both sides (Devlin et al., 2019). In decoder-style models such as GPT-like systems, attention is masked so that token t can attend only to positions $\leq t$, thereby supporting autoregressive next-token prediction (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023). If $D^{(\ell-1)} \in \mathbb{R}^{m\times d}$ denotes decoder states entering layer ℓ, masked self-attention can be written as

$$\mathrm{MaskedAttn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top + M}{\sqrt{d_k}}\right)V, \tag{2.10}$$

where $M \in \mathbb{R}^{m\times m}$ is a causal mask with

$$M_{ij} = \begin{cases} 0, & j \leq i, \\ -\infty, & j > i. \end{cases} \tag{2.11}$$

This mask prevents position i from attending to future positions $j > i$, ensuring that generation depends only on the already produced prefix.

In the original encoder-decoder transformer, the decoder contains both masked self-attention and a cross-attention sublayer that conditions on encoder outputs. If $E \in \mathbb{R}^{n\times d}$ denotes encoder representations and $D^{(\ell-1)} \in \mathbb{R}^{m\times d}$ denotes decoder states entering layer ℓ, then cross-attention can be written as

$$\mathrm{CrossAttn}\big(D^{(\ell-1)}, E\big) = \mathrm{Attention}\big(D^{(\ell-1)}W_Q,\ EW_K,\ EW_V\big). \tag{2.12}$$

This means that the decoder uses its current states as queries while treating the encoded source sequence as keys and values. In machine translation, for example, the decoder can use cross-attention to query the source sentence when generating the next target word. The same general mechanism later becomes important in multimodal settings, where a decoder may instead query visual tokens or projected image representations. Figure 2.2 places these components within the original encoder-decoder architecture and shows where masked self-attention and encoder-decoder attention occur.

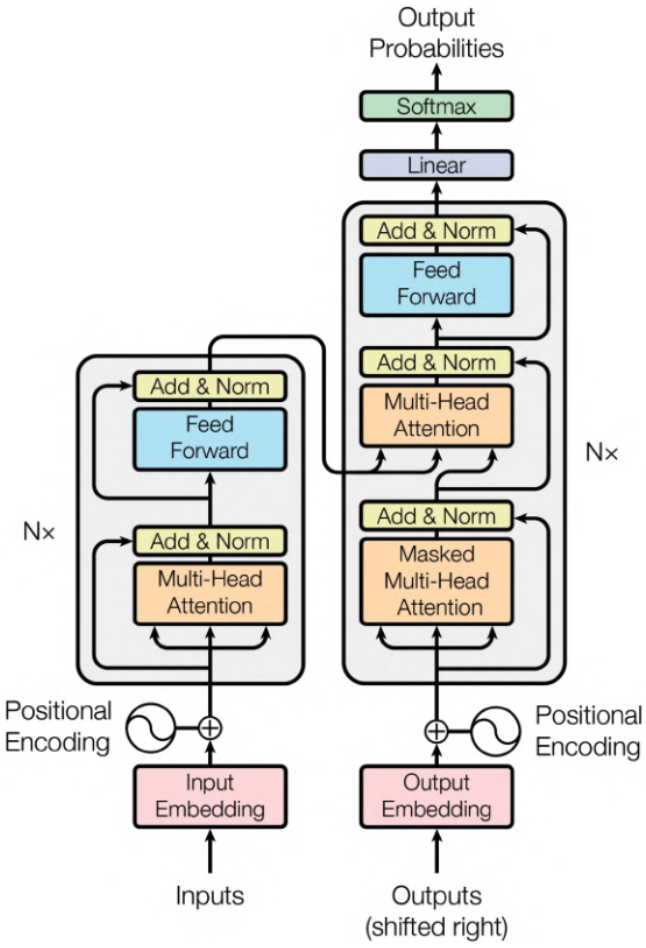


Figure 2.2: Schematic overview of the original transformer architecture, showing the encoder stack, decoder stack, positional encoding, masked self-attention, and encoder-decoder attention. Adapted from (Vaswani et al., 2017).

Within this dissertation, two properties of transformers are especially important. First, their layered hidden representations make it possible to study where particular kinds of information are encoded and where targeted interventions may be most effective (Geva et al., 2021; D Dai et al., 2022). Second, the same sequence-modeling machinery extends naturally to multimodal settings once visual inputs are converted into token-like embeddings, which is why transformer-based architectures also underpin large vision-language models (J Li et al., 2022; Haotian Liu et al., 2023). For the chapters that follow, transformers therefore serve not merely as background architecture, but as the common computational substrate for multimodal alignment, efficient adaptation, autoregressive generation, and representation-level steering.

2.1.2 Vision-Language Modeling, Fusion, and Alignment

Vision-language modeling studies how visual and textual signals are represented, aligned, and jointly processed within a shared computational framework. Let an image input be denoted by I, and let a tokenized text input be denoted by $x = (x_1, \ldots, x_n)$, where x_t is the token at position t. In a vision-language model, the image is converted into

a sequence of visual token embeddings $v = (v_1, \dots, v_m)$ with $v_j \in \mathbb{R}^d$, while the text tokens are mapped to embeddings in a compatible representational space. Here, n and m denote the text and visual sequence lengths, respectively. Depending on the architecture, the visual sequence may be derived from detector-proposed object regions encoded separately (Girshick, 2015; Jiasen Lu et al., 2019), from fixed-size image patches projected into vectors (Dosovitskiy et al., 2021; Kim et al., 2021), or from a learned resampling mechanism that compresses a larger set of perceptual features into a smaller token set (Alayrac et al., 2022). In all cases, the goal is to convert visual input into token-like representations that can interact with language representations inside a common modeling pipeline. For simplicity, the discussion here begins with images, but videos can be treated analogously as temporally extended visual token sequences. The resulting systems support tasks such as cross-modal retrieval, visual question answering, caption generation, and multimodal dialogue (Karpathy and Fei-Fei, 2015; Faghri et al., 2018; Jiasen Lu et al., 2019; Chen et al., 2020; J Li et al., 2022).

Within this setting, it is useful to distinguish *alignment* from *fusion*. Alignment concerns whether visual and textual inputs are mapped into representational spaces in which corresponding image-text pairs are close and mismatched pairs are far apart (Radford et al., 2021; Faghri et al., 2018). Fusion concerns the architectural mechanisms through which token-level or feature-level information from the two modalities interacts during processing (Jiasen Lu et al., 2019; Chen et al., 2020). A retrieval task illustrates the first notion: the model must decide whether an image and a caption express the same overall content. A visual question answering task illustrates the second: if the question asks "What is the dog holding?", the model must combine the linguistic query with localized visual evidence rather than relying on global similarity alone. In practice, strong vision-language systems typically rely on both, because alignment supports coarse cross-modal correspondence while fusion enables fine-grained conditional reasoning.

One family of vision-language models therefore emphasizes fusion within the architecture. Dual-stream models maintain partially separate visual and textual representations and exchange information through cross-attention or co-attention modules, allowing each modality to condition on the other while preserving some modality-specific structure (Jiasen Lu et al., 2019). Single-stream models instead combine visual and textual embeddings within a shared transformer, producing more tightly integrated multimodal representations (Chen et al., 2020; X Li et al., 2020; Kim et al., 2021). A related line of work explicitly separates alignment and fusion stages, first encouraging modality-compatible representations and then refining them through deeper interaction (J Li et al., 2021; Bao et al., 2022). Patch-based approaches further simplify this pipeline by reducing dependence on region detectors and moving vision-language pretraining closer to end-to-end transformer modeling (Kim et al., 2021; W Wang et al., 2023). Across these designs, the central architectural question is how much modality-specific structure should be preserved and at what stage cross-modal interaction should occur.

A second family of approaches emphasizes global alignment through contrastive learning. Consider a minibatch of paired image-text examples $\{(I_i, T_i)\}_{i=1}^{B}$, where B is the minibatch size. Let

$$z_i^v = f_v(I_i), \qquad z_i^x = f_x(T_i), \tag{2.13}$$

denote the corresponding normalized global visual and textual embeddings in $\mathbb{R}^d$ produced by modality-specific encoders. These global embeddings are typically obtained by

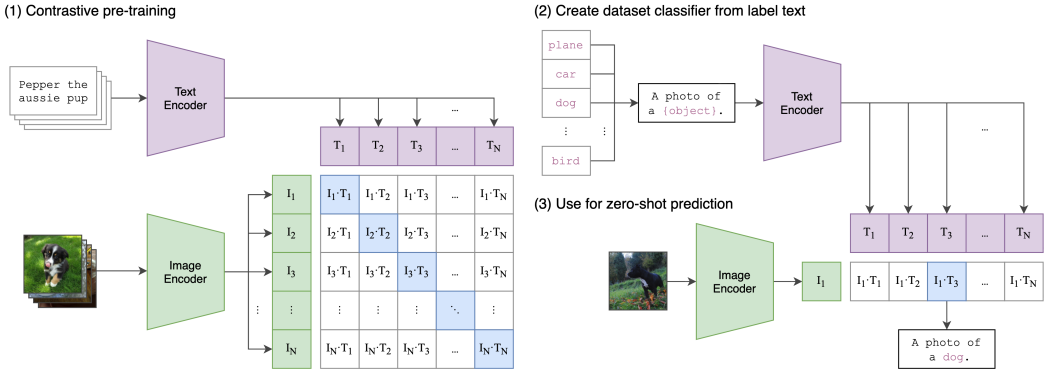


Figure 2.3: Illustration of contrastive vision-language pretraining and zero-shot transfer in CLIP. Image and text encoders are trained to align matched image-text pairs in a shared embedding space, after which textual label prompts can be used for zero-shot prediction. Adapted from (Radford et al., 2021).

pooling or projecting modality-specific token representations into a shared contrastive space. A common similarity score is

$$s_{ij} = \frac{(z_i^v)^\top z_j^x}{\tau}, \tag{2.14}$$

where $\tau > 0$ is a temperature parameter. A standard symmetric contrastive objective is

$$\mathcal{L}_{\text{align}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ij})} + \log \frac{\exp(s_{ii})}{\sum_{j=1}^{B} \exp(s_{ji})} \right]. \tag{2.15}$$

Objectives of this type encourage matched image-text pairs to be close in a shared embedding space while pushing mismatched pairs apart (Radford et al., 2021; Faghri et al., 2018). In practice, the off-diagonal pairs within the minibatch act as implicit negatives, and the symmetric loss jointly supports image-to-text and text-to-image retrieval. This form of pretraining is highly effective for learning broad semantic correspondence, but by itself it mainly provides global alignment. Additional fusion, supervision, or generation-oriented training is typically needed when token-level grounding or context-sensitive interpretation must be modeled explicitly.

Figure 2.3 illustrates a canonical contrastive setup in which matched image-text pairs are aligned in a shared embedding space and the resulting representations support zero-shot transfer.

A useful historical progression runs from alignment-oriented models to systems that combine alignment, fusion, and open-ended generation. Architectures such as BLIP (J Li et al., 2022) and CoCa (J Yu et al., 2022) integrate contrastive learning with captioning or language-modeling losses, thereby supporting both discriminative and generative multimodal behavior. Flamingo (Alayrac et al., 2022) provides a particularly clear bridge between these stages. In Flamingo, a frozen vision encoder first extracts perceptual features, a Perceiver Resampler compresses them into a fixed set of visual tokens, and gated cross-attention layers inject those tokens into a largely frozen autoregressive language model. More specifically, the vision backbone and language backbone remain frozen, while the Perceiver Resampler and gated cross-attention layers are trained to mediate between vision and language. This design is conceptually important because it

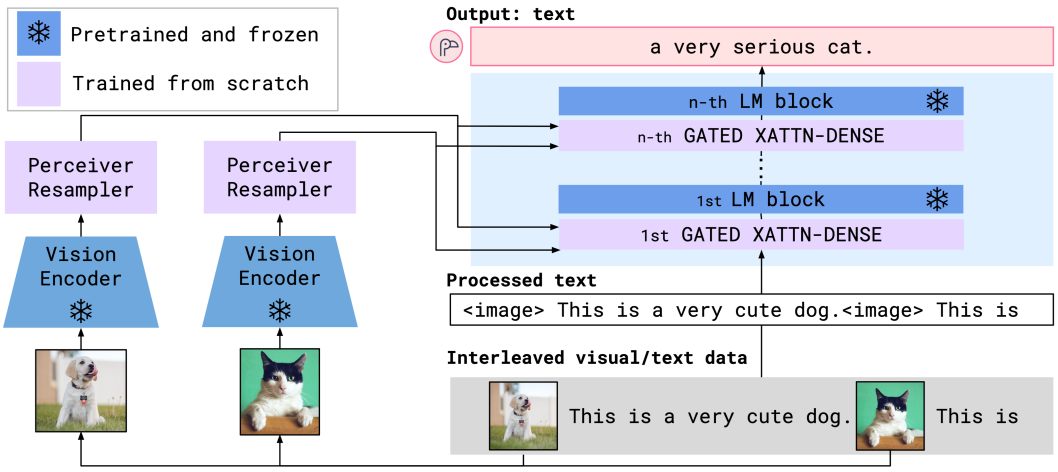


Figure 2.4: Illustration of Flamingo-style multimodal generation. A frozen vision encoder extracts image features, a Perceiver Resampler compresses them into a fixed set of visual tokens, and gated cross-attention layers inject those tokens into a largely frozen autoregressive language model. Adapted from (Alayrac et al., 2022).

shows how visual information can condition generation without requiring the entire language backbone to be retrained from scratch, and it supports interleaved image-text prompting in which the model repeatedly queries compact visual tokens while generating text.

Figure 2.4 illustrates this intermediate design, in which visual inputs are converted into a compact token set and then connected to a text-generating language model through gated cross-attention.

Contemporary open large vision-language models extend this generative pattern further and make the underlying architecture more explicit. A representative contemporary example is Qwen2.5-VL (Bai, Chen, et al., 2025), which combines a vision encoder with a decoder-style language model while supporting both images and videos. Its architecture further highlights several characteristics that have become increasingly important in modern multimodal assistants, including native-resolution visual input, temporal sampling for video, and variable numbers of visual tokens under token compression and budgeting strategies. Rather than treating visual input as a single global vector, such systems process perceptual content as structured token sequences that can be interleaved with text and consumed by the language model during generation. This design is especially relevant for the present dissertation because it more closely resembles the class of open multimodal assistants used in later chapters than earlier retrieval-oriented or encoder-only architectures.

Figure 2.5 highlights these modern architectural characteristics. In particular, it shows that images and videos are converted into token sequences by a vision encoder, optionally compressed or sampled according to input size and temporal structure, and then passed to a decoder-centric language model that performs open-ended multimodal generation.

Beyond this stage, a broader shift toward *native multimodality* has become increasingly visible. Newer systems such as Qwen3-VL (Bai, Cai, et al., 2025), InternVL3 (J Zhu et al., 2025), and Qwen3.5 (Qwen Team, 2026) place stronger emphasis on multimodal capability from earlier stages of model design and pretraining rather than treating vision

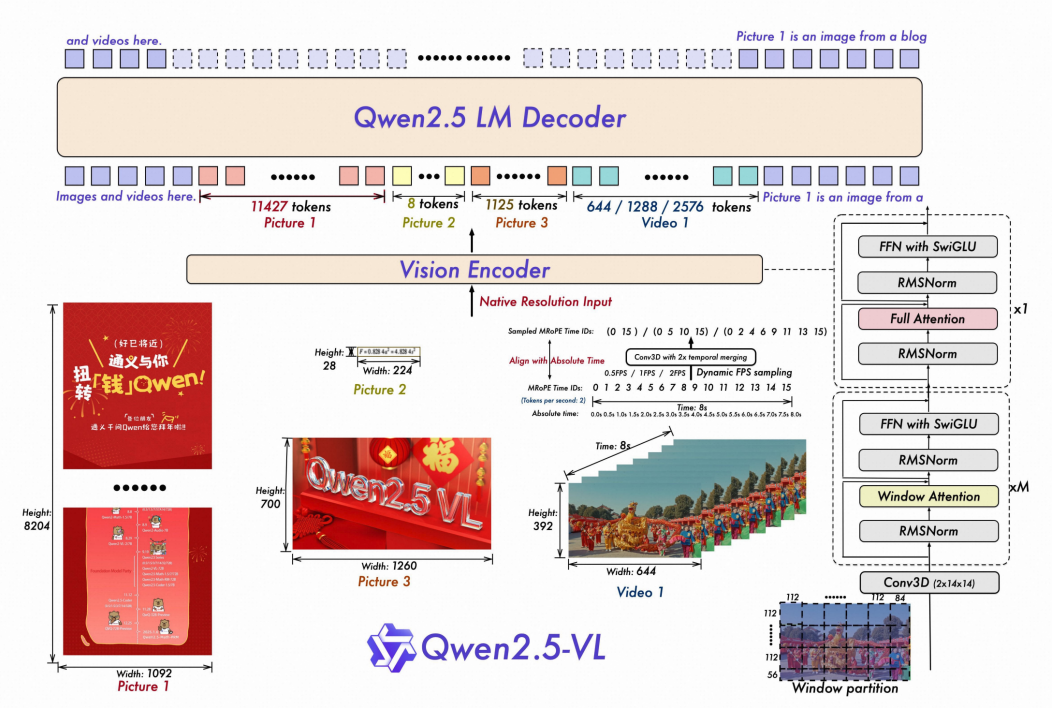


Figure 2.5: Representative architecture of a modern open large vision-language model, illustrated with Qwen2.5-VL. The model combines a vision encoder with a decoder-style language model, supports both images and videos, and allocates variable numbers of visual tokens through native-resolution processing, temporal sampling, and token compression. Adapted from (Bai, Chen, et al., 2025).

solely as a late-stage extension of a text-only model. These developments substantially expand capability, but they also make vision-language modeling more sensitive to failures of grounding, faithfulness, hallucination, and controllability at inference time.

For the later empirical chapters, three distinctions are especially important. First, multimodal grounding depends on whether alignment mechanisms capture contextually relevant relations between language and perception, rather than only coarse semantic similarity. Second, efficient adaptation must preserve the cross-modal structure established during pretraining when models are specialized to new domains or tasks. Third, generative large vision-language models make it necessary to distinguish between jointly aligned internal representations and outputs that remain faithful to the conditioning evidence during decoding. These distinctions motivate the later discussion of context-sensitive grounding, stable multimodal adaptation, and trustworthy multimodal generation.

2.1.3 Adaptation Regimes: Zero-Shot, Fine-Tuning, and Parameter-Efficient Tuning

After large-scale pretraining, foundation models are typically specialized to downstream tasks or domains under several adaptation regimes. Let f_{θ_0} denote a pretrained model with parameters θ_0, let c denote the task-conditioning context, and let $\mathcal{D}$ denote downstream training data when such data are available. In this subsection, adaptation is understood in a broad sense to include both conditioning-only deployment regimes, in

which only the conditioning context changes at deployment time and no post-pretraining parameter updates are made, and parameter-updating regimes after pretraining. At a high level, these methods differ in what is allowed to change after pretraining:

$$\begin{aligned} \text{conditioning only: } & \theta = \theta_0, \\ \text{PEFT: } & \phi^\star = \arg\min_{\phi} \mathcal{L}(\mathcal{D}; \theta_0, \phi), \\ \text{full FT: } & \theta^\star = \arg\min_{\theta} \mathcal{L}(\mathcal{D}; \theta), \end{aligned} \tag{2.16}$$

where ϕ denotes a relatively small set of trainable task-specific parameters with $|\phi| \ll |\theta_0|$, and $\mathcal{L}$ is the downstream objective. These regimes therefore differ not only in computational cost, but also in the locus and extent of representational change.

Zero-shot and few-shot in-context learning belong to the first category because they change only the deployment context and do not update model parameters at all (Brown et al., 2020; Bommasani et al., 2021; Achiam et al., 2023). Instead, task behavior is elicited through natural-language instructions, demonstrations, or examples included directly in the input context. In a language model, this may take the form of an instruction together with a small number of input-output pairs. In a multimodal assistant, it may take the form of an image-conditioned instruction such as "Describe the abnormality in this image" or "Answer the question using the visual evidence only," again without changing the pretrained weights (Haotian Liu et al., 2023; J Li et al., 2022). The main advantage of such regimes is that they preserve the pretrained model exactly, but their effectiveness depends heavily on prompt formulation, task format, and the extent to which the relevant capability has already been acquired during pretraining.

At the other extreme, full fine-tuning updates all parameters with respect to a downstream objective. This strategy can yield strong task-specific performance, especially when sufficient labeled data are available, but it is computationally expensive and difficult to scale across many tasks or domains. More importantly for the present dissertation, unconstrained parameter updates may distort semantic relationships learned during pretraining, including cross-modal correspondences and other structural regularities (H Li et al., 2024). These are precisely the kinds of structure that later chapters seek to preserve. For example, a multimodal model pretrained on broad image-text corpora may later be adapted to a specialized task such as domain-specific visual question answering or report generation. Full fine-tuning can improve downstream accuracy, but it may also reshape the internal geometry of the model in ways that weaken previously useful alignment structure. These concerns motivate parameter-efficient tuning methods that seek useful specialization while limiting disruption to the pretrained backbone.

A first family of parameter-efficient methods learns *continuous prompts* rather than updating the full model. Prompt tuning (Lester et al., 2021) learns trainable soft embeddings at the input layer, P-Tuning (X Liu et al., 2022) generalizes this idea with more flexible prompt parameterizations, and prefix tuning (Li and Liang, 2021) extends it further by injecting trainable prefix states into the attention computation of multiple layers (Figure 2.6). For a single attention head at layer ℓ, a simplified prefix-tuning formulation can be written as

$$\text{PrefixAttn}^{(\ell)}(Q, K, V) = \text{Attention}\left(Q, \begin{bmatrix} P_K^{(\ell)} \\ K \end{bmatrix}, \begin{bmatrix} P_V^{(\ell)} \\ V \end{bmatrix}\right), \tag{2.17}$$

where $Q \in \mathbb{R}^{n \times d_k}$, $K \in \mathbb{R}^{n \times d_k}$, and $V \in \mathbb{R}^{n \times d_v}$ are the usual query, key, and value matrices, while $P_K^{(\ell)} \in \mathbb{R}^{m_p \times d_k}$ and $P_V^{(\ell)} \in \mathbb{R}^{m_p \times d_v}$ are trainable prefix states and m_p is the prefix

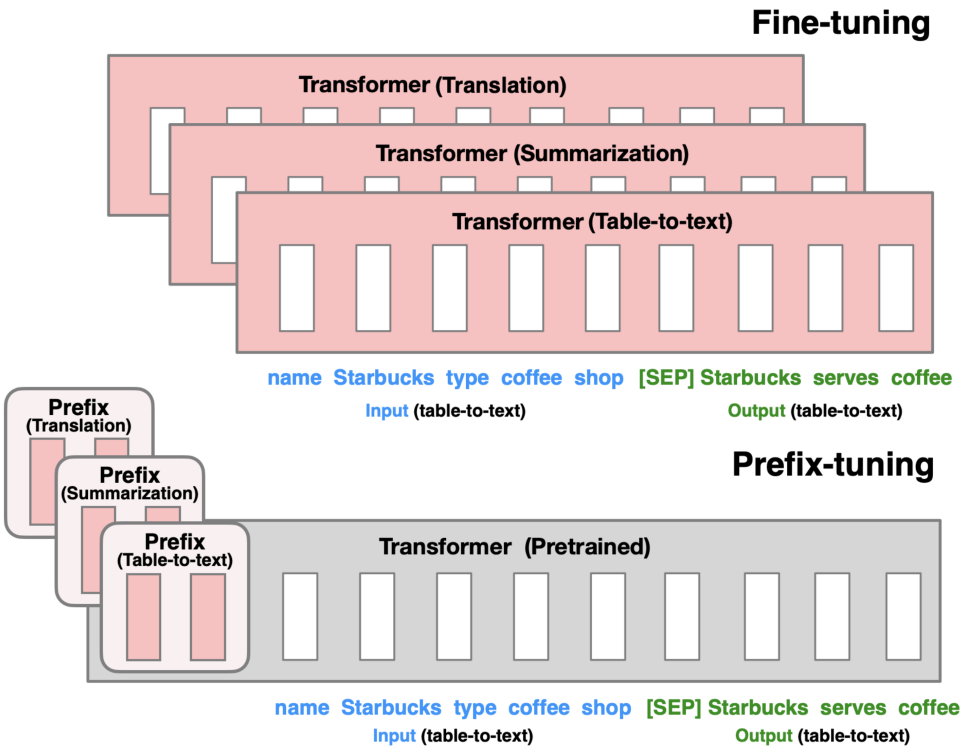


Figure 2.6: Illustration of prefix-tuning as a parameter-efficient alternative to full fine-tuning. Task-specific continuous prefix parameters are introduced while the pretrained transformer backbone remains frozen. Adapted from (Li and Liang, 2021).

length. The pretrained backbone remains frozen, but the learned prefix parameters alter how attention is computed for the downstream task. This distinguishes continuous prompting from ordinary natural-language prompting: the prompt is no longer only text supplied by the user, but also a set of trainable vectors optimized by gradient descent. In multimodal settings, an analogous idea appears on the visual side in visual prompt tuning, where trainable prompt parameters are attached to the vision backbone rather than to the text input (Jia et al., 2022).

A second family inserts small trainable modules into the frozen network. Adapter-based approaches (Houlsby et al., 2019) place lightweight bottleneck layers inside each transformer block so that task-specific behavior can be learned without modifying most of the original parameters (Figure 2.7). If $h \in \mathbb{R}^d$ denotes a hidden representation at a given layer and position, a standard bottleneck adapter can be written as

$$\text{Adapter}(h) = h + W_{\text{up}}\, \phi(W_{\text{down}} h)\,, \tag{2.18}$$

where $W_{\text{down}} \in \mathbb{R}^{r \times d}$ projects the representation into a lower-dimensional bottleneck with $r \ll d$, $W_{\text{up}} \in \mathbb{R}^{d \times r}$ projects it back to the original space, and ϕ is a nonlinear activation function. The original layer computation remains intact, while the adapter adds a small residual transformation around it. In the Houlsby configuration, adapters are inserted after both the multi-head attention and feed-forward sublayers within each transformer block (Houlsby et al., 2019). In multimodal settings, related ideas have been extended through methods such as VL-Adapter (Sung et al., 2022) and SparseAdapter (S He et al., 2022), which may target the vision encoder, the language model, or cross-modal interaction layers.

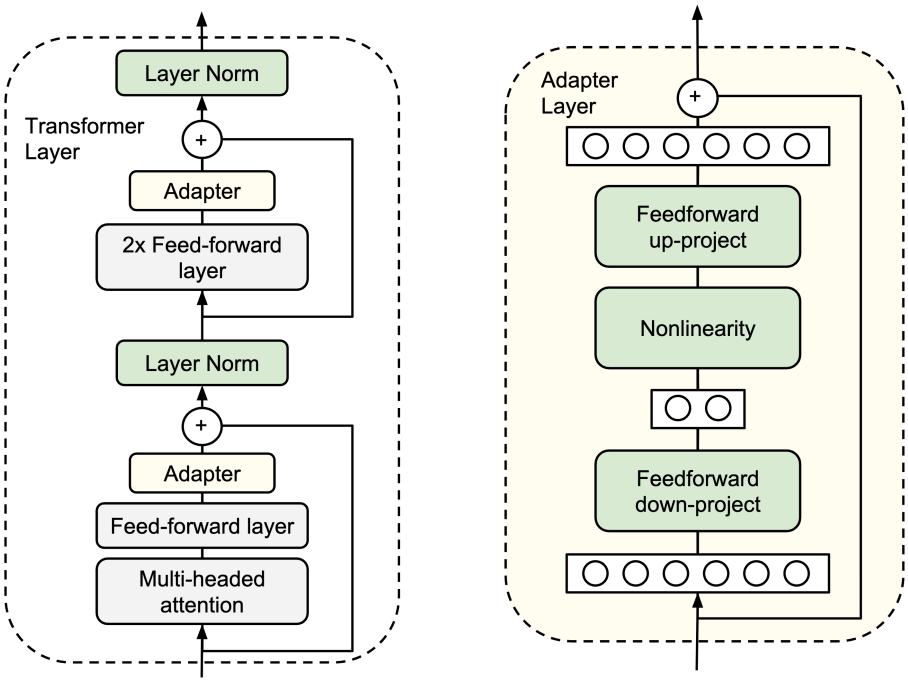


Figure 2.7: Illustration of adapter-based parameter-efficient tuning. Small bottleneck modules are inserted into transformer blocks, allowing task-specific adaptation while leaving the original backbone parameters largely frozen. Adapted from (Houlsby et al., 2019).

A third family constrains adaptation at the level of existing weight matrices rather than by inserting new modules. Low-rank adaptation methods such as LoRA (Hu et al., 2022) update the weight matrix of a linear transformation $W \in \mathbb{R}^{d\times k}$ by adding a trainable low-rank increment:

$$W' = W + \Delta W, \qquad \Delta W = \frac{\alpha}{r} BA, \tag{2.19}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, α is a scaling parameter, and the rank r is chosen such that $r \ll \min(d, k)$. Here, W denotes the weight matrix of a pretrained linear layer, so LoRA does not replace the original transformation but learns a constrained low-rank update on top of it. In contrast to full fine-tuning, which learns an unconstrained dense update to the pretrained parameters, LoRA restricts the update to a low-rank subspace. In practice, LoRA is often applied to projection matrices in attention or feed-forward layers, which allows adaptation to act directly on the transformations already used by the model while keeping the number of trainable parameters small.

This low-rank parameterization is visualized in Figure 2.8, which highlights how task-specific updates can be introduced without directly updating the full pretrained weight matrix.

For multimodal foundation models, these adaptation choices matter because trainable parameters can be attached to different parts of the system: the vision encoder, the language model, or the layers that mediate cross-modal interaction. The central issue is therefore not only whether adaptation is efficient, but also whether it preserves the representational integrity of the pretrained model. Recent work has shown that even lightweight tuning can alter internal geometry, weaken previously learned correspondences, or introduce forms of forgetting that are not visible from downstream accuracy alone (H Li et al., 2024; Shihab et al., 2026; K Yao et al., 2026). In the later chapters of this

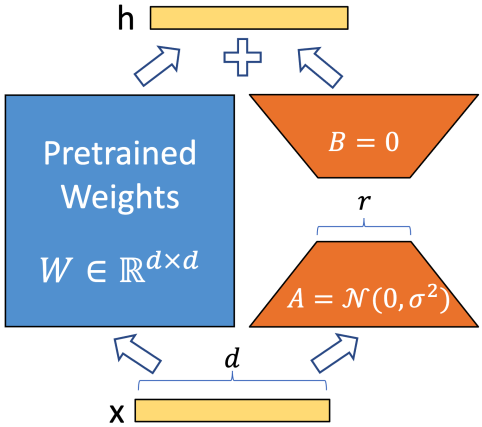


Figure 2.8: Illustration of low-rank adaptation in LoRA. Instead of updating the full pretrained weight matrix W, adaptation is represented as a low-rank increment parameterized by matrices A and B, reducing the number of trainable parameters. Adapted from (Hu et al., 2022).

dissertation, adaptation is therefore treated not merely as a problem of computational efficiency, but as a constrained transformation of pretrained representations whose inter-modal alignment and intra-modal semantic structure should ideally remain usable after specialization.

2.1.4 Inference-Time Generation, Decoding, and Steering

This subsection focuses on autoregressive decoders, because they underlie the large language models and generative vision–language models most relevant to the later chapters of this dissertation. Let c denote the conditioning context, which may include textual prompts, task instructions, and, in multimodal settings, visual tokens or projected visual embeddings derived from an input image. If the generated output sequence is $y = (y_1, \dots, y_T)$, then autoregressive generation factorizes the conditional probability of the sequence as

$$p(y \mid c) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, c), \tag{2.20}$$

where $y_{<t} = (y_1, \dots, y_{t-1})$ denotes the previously generated tokens. Generation therefore proceeds step by step: at each time step, the model predicts a distribution over the next token conditioned on the context and the partial output produced so far.

To make this prediction more explicit, let $\mathcal{V}$ denote the output vocabulary, let $h_t^{(L)} \in \mathbb{R}^d$ denote the top-layer hidden state at time step t, and let $z_t \in \mathbb{R}^{|\mathcal{V}|}$ denote the corresponding vector of token scores or logits. A common parameterization computes

$$z_t = W_{\mathrm{LM}} h_t^{(L)} + b, \qquad p(y_t = w \mid y_{<t}, c) = \mathrm{softmax}(z_t)_w, \tag{2.21}$$

where $W_{\mathrm{LM}} \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the output projection matrix, $b \in \mathbb{R}^{|\mathcal{V}|}$ is a bias term, and $w \in \mathcal{V}$ is a candidate next token. This relation is important because it links internal hidden representations to observable generation behavior: hidden states determine logits, logits determine probabilities, and decoding selects actual continuations from those probabilities.

Decoding refers to the rule used to transform the next-token distribution into concrete output tokens. Under greedy decoding (Bahdanau et al., 2016), the selected token at time step t is

$$\hat{y}_t^{\text{greedy}} = \arg\max_{w \in \mathcal{V}} p(y_t = w \mid y_{<t}, c). \tag{2.22}$$

Beam search (Bahdanau et al., 2016) instead keeps several partial hypotheses in parallel and compares their cumulative log-probability or related sequence-level scores, thereby approximating sequence-level search rather than making a purely local token-level choice at each step. Stochastic methods such as sampling draw tokens from the predicted distribution, often after applying temperature scaling or top-k truncation (Radford et al., 2019), or nucleus (top-p) filtering (Holtzman et al., 2020). A temperature-scaled distribution can be written as

$$p_\tau(y_t = w \mid y_{<t}, c) = \text{softmax}(z_t/\tau)_w, \tag{2.23}$$

where $\tau < 1$ sharpens the distribution, $\tau > 1$ flattens it, and $\tau = 1$ recovers the original distribution. These choices substantially affect fluency, diversity, and factual reliability. Even when model parameters remain fixed, the generated behavior can change markedly depending on how the output distribution is searched, truncated, or sampled. Decoding is therefore not merely a procedural detail, but part of the effective behavior of a generative model.

One broad family of inference-time control methods operates directly in the *output space* by modifying token scores or output distributions at inference time, without additional parameter updates to the base generator. This includes methods based on expert and anti-expert combinations such as DExperts (A Liu et al., 2021), contrastive decoding (XL Li et al., 2023), layer-contrastive decoding such as DoLa (Chuang et al., 2024), and multimodal contrastive strategies (Leng et al., 2024) that compare predictions under different visual or contextual conditions. Among these methods, contrastive decoding provides a particularly clear worked example. A simplified contrastive decoding score can be written as

$$s_{\text{CD}}(w) = \log p_{\text{exp}}(w \mid y_{<t}, c) - \log p_{\text{ama}}(w \mid y_{<t}, c), \tag{2.24}$$

where p_{exp} and p_{ama} denote the next-token distributions of a stronger expert model and a weaker amateur model, respectively. The two distributions are compared over the same candidate vocabulary at the same decoding step. Instead of selecting tokens purely by expert probability, contrastive decoding favors tokens that are preferred by the stronger model but not by the weaker one. This simplified expression captures the core contrastive idea, abstracting away additional plausibility constraints used in specific implementations. Figure 2.9 therefore focuses on this contrastive expert-amateur setup rather than on the full variety of output-space control methods.

This family of methods is especially relevant in multimodal generation because unsupported continuations often arise when internal language priors dominate perceptual evidence during decoding. For example, an image-conditioned caption may mention an unsupported object such as a frisbee even when no such object is present in the image. Output-space control can be used to reduce the probability of such unsupported continuations without retraining the model. In this sense, decoding-time control is directly connected to multimodal faithfulness: it changes what the model says by altering how its next-token distribution is queried or reweighted.

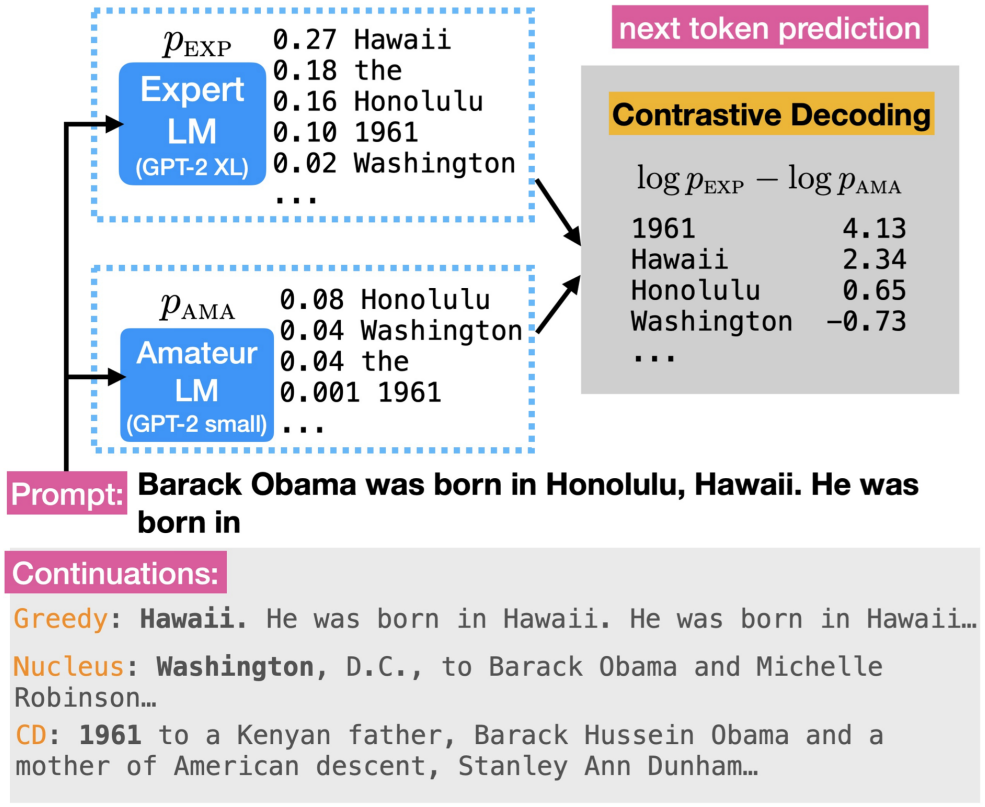


Figure 2.9: Illustration of contrastive decoding using an expert and an amateur language model. At each decoding step, token scores are adjusted by contrasting the expert distribution with a weaker reference distribution, encouraging continuations that are favored by the stronger model but not by the weaker one. Adapted from (XL Li et al., 2023).

A second family of methods operates in the *representation space* by intervening directly in hidden states during inference. Let $h_t^{(\ell)} \in \mathbb{R}^{d_\ell}$ denote the hidden representation at layer ℓ and time step t. A generic additive hidden-state intervention can be written as

$$\widetilde{h}_t^{(\ell)} = h_t^{(\ell)} + \lambda d^{(\ell)}, \tag{2.25}$$

where $d^{(\ell)} \in \mathbb{R}^{d_\ell}$ is an intervention direction and $\lambda \in \mathbb{R}$ controls the strength of the intervention. The direction $d^{(\ell)}$ can be interpreted as a behavior-relevant direction in hidden-state space, estimated for example from examples, probes, or contrasts between target behaviors. The intervention is applied at a chosen layer ℓ and time step t, after which generation continues from the modified state. If $\ell < L$, the modified state propagates through the remaining decoder layers before reaching the output head. Variants of this idea include activation steering (Su et al., 2025), layerwise interventions, and methods such as CogSteer (**Wang**, Pan, Ding, et al., 2025), which bias generation toward or away from particular behaviors by modifying the model's internal trajectory. Because the hidden state ultimately determines the logits in Equation 2.21, representation-space interventions provide a more localized route for changing output behavior than full fine-tuning.

For the generation- and control-oriented studies that follow, three aspects of inference-time behavior are especially important. First, trustworthy multimodal generation depends not only on what a model has learned during pretraining, but also on whether

decoding remains faithful to the conditioning evidence at inference time. Second, output-space control provides a mechanism for runtime control without additional parameter updates to the base generator when retraining is undesirable or infeasible. Third, representation-space control makes it possible to connect interpretability to intervention by identifying and modifying the internal states most relevant to downstream behavior. In this dissertation, inference-time generation is therefore treated not merely as a post-processing stage, but as a central site at which faithfulness, controllability, and internal representation structure become operational.

2.1.5 Human Cognitive Signals and Gaze as External Guidance

Human cognitive signals can provide external evidence about how information is processed during language understanding. Such signals may arise from eye movements, reading times, or other behavioral and physiological measurements. In this dissertation, the focus is on gaze during reading, because eye-tracking provides a temporally fine-grained and linguistically aligned record of how readers allocate attention over text (Rayner, 1998; Rayner and Pollatsek, 2016). In psycholinguistics, gaze-based evidence has long been used to study lexical access, syntactic ambiguity, semantic integration, and processing difficulty. Although eye movements do not provide a direct or exhaustive representation of cognition, they offer a behaviorally grounded window into how comprehension unfolds incrementally over a sentence.

At the level of raw observation, an eye-tracking record consists of temporally ordered eye-movement events. A *fixation* is a relatively stable period during which the eyes remain on a location and visual information is acquired, whereas a *saccade* is the rapid movement between two fixations (Rayner, 1998; Rayner and Pollatsek, 2016). During reading, a *regression* refers to a backward eye movement to earlier text, often indicating reanalysis, integration difficulty, or uncertainty under the current context. To derive word-level measures, fixations are typically assigned to predefined word-level areas of interest, from which word-level statistics are aggregated. These raw events therefore form the basis from which more interpretable reading measures are constructed. Figure 2.10 illustrates these basic components, including fixations, saccades, regressions, and the distinction between first-pass and later reading behavior.

From these raw events, researchers derive word-level gaze measures that can be aligned to text. Common examples include *first fixation duration* (FFD), the duration of the first fixation landing on a word; *gaze duration* (GD), which for a single word region is often operationalized as *first-pass reading time*, namely the total fixation time spent on a word during the first-pass reading sequence before the eyes leave it; and *total reading time* (TRT), the overall time spent on a word including later returns (Rayner, 1998; Rayner and Pollatsek, 2016). Regressions to a word or region can also be summarized as count- or duration-based indicators. These measures are informative because reading is incremental: readers do not process all words uniformly, but allocate overt visual attention selectively as comprehension unfolds. Longer fixations, longer gaze durations, and more frequent regressions are often associated with increased ambiguity, integration demands, or processing difficulty, whereas shorter and more stable patterns can indicate easier interpretation under the current context.

For computational modeling, these measurements can be represented as word-aligned feature vectors over a sentence. Let the word sequence be denoted by $w = (w_1, \dots, w_n)$, where w_i is the i-th word. Let $g_i \in \mathbb{R}^m$ denote the reader-aggregated or

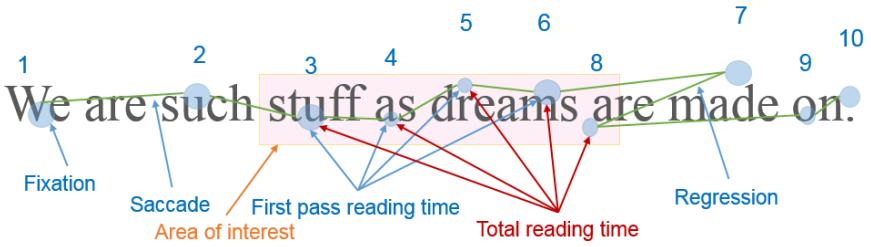


Figure 2.10: Illustration of common eye-tracking measures during reading, including fixations, saccades, regressions, first-pass reading time, and total reading time. Such measures provide word- or region-level evidence about how processing effort is distributed over text. Adapted from (Eckstein et al., 2019).

normalized gaze feature vector aligned to word w_i, where the m dimensions may include measures such as FFD, GD, TRT, or regression-based statistics. The gaze record for the full sequence can then be written as

$$G = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix} \in \mathbb{R}^{n \times m}. \tag{2.26}$$

Words that are skipped during reading require explicit preprocessing conventions, such as zero-valued features, masked entries, or other task-specific handling.

A further practical challenge arises because gaze annotations are usually word-level, whereas transformer models often use subword tokenization. Let $h_j^{(\ell)}$ denote the subword-level hidden state at position j and layer ℓ, and let $A(i)$ denote the set of subword positions corresponding to word w_i. A word-aligned hidden representation can then be obtained by aggregating the subword states associated with that word:

$$\bar{h}_i^{(\ell)} = \mathrm{Pool}\left(\{h_j^{(\ell)} \,:\, j \in A(i)\}\right), \tag{2.27}$$

where Pool may denote mean pooling, sum pooling, or a first-subword selection rule. Collecting these word-level representations gives

$$\bar{H}^{(\ell)} = \begin{bmatrix} \bar{h}_1^{(\ell)} \\ \vdots \\ \bar{h}_n^{(\ell)} \end{bmatrix} \in \mathbb{R}^{n \times d}. \tag{2.28}$$

This alignment step is important because it determines how cognitively grounded human signals can be related to model-internal representations in a formally consistent way.

Word-aligned eye-tracking corpora have made such comparisons increasingly feasible. The Provo Corpus provides natural reading data together with word-level eye-tracking and contextual predictability information, making it useful for studying how expectation and contextual variation affect lexical processing (Luke and Christianson, 2018). GECO provides word-level eye-tracking data over extended reading and has been useful for studying sentence- and discourse-level reading behavior (Colman et al., 2022). ZuCo combines reading data with eye-tracking and additional neural recordings, enabling richer multimodal analyses of language processing (Hollenstein et al., 2020). These resources do not merely add another layer of annotation; they provide an external

behavioral reference for testing whether model representations capture distinctions that matter for human processing.

Within this dissertation, the role of gaze is not to replace conventional supervision, but to provide cognitively grounded external guidance for model analysis and intervention. Once gaze features are aligned to words, the gaze matrix G can be compared with word-aligned hidden representations $\bar{H}^{(\ell)}$ across layers to ask which layers align most systematically with human processing signals. If particular layers or internal representations track gaze-derived patterns more consistently, this can help identify where semantically or behaviorally relevant information is concentrated. In this way, gaze supports a two-step bridge from analysis to control: it helps indicate where to look inside a model, and it can help motivate where targeted intervention may be most meaningful. This perspective is especially relevant for the later chapters on interpretability and steering, where human-centered processing signals are used not only to describe model behavior, but also to inform representation-level intervention.

2.2 From Foundations to Research Problems

The preceding section has established the technical and methodological foundations needed for the remainder of the dissertation, spanning core modeling mechanisms such as transformer representations, multimodal alignment, adaptation, and inference-time control, as well as human-centered external signals such as gaze during reading. Together, these foundations clarify how contemporary foundation models represent information, adapt to new tasks, and generate outputs. They do not, however, by themselves resolve where such models remain insufficiently grounded, difficult to adapt without representational drift, prone to unfaithful generation, or hard to control in a targeted and meaning-preserving way. The next step is therefore to move from foundations to the specific research problems that motivate the dissertation.

Against this background, the remainder of this chapter is organized not as a comprehensive catalog of prior work, but as a structured formulation of the dissertation's core problem areas. The following subsections identify five recurring tensions: grounding without sufficient contextual sensitivity, adaptation without preserved representation integrity, generation without adequate faithfulness, safety without pragmatic preservation, and interpretability without operational control. Taken together, these tensions define the research space in which the later empirical chapters are situated, linking the technical foundations above to the problem-driven structure of the later empirical chapters.

2.2.1 Limitations of Grounding Without Context

A central objective of vision–language research has been to enable models to associate linguistic expressions with visual content through large-scale multimodal pretraining (Chen et al., 2020; X Li et al., 2020; Radford et al., 2021; J Li et al., 2022; J Yu et al., 2022). Approaches based on contrastive learning and image–text pairing have demonstrated that shared embedding spaces can capture broad semantic correspondences, supporting tasks such as cross-modal retrieval, caption generation, and visual question answering (Karpathy and Fei-Fei, 2015; Faghri et al., 2018; Plummer et al., 2015; Hudson and Manning, 2019). These methods have played a foundational role in establishing

multimodal foundation models, showing that large datasets and aligned objectives can produce transferable representations across modalities.

Despite these advances, much of the existing work emphasizes global semantic alignment—that is, learning whether an image and a text are generally related—rather than examining how visual information supports interpretation within a specific linguistic context for human-centered understanding (Harnad, 1990; Radford et al., 2021; J Li et al., 2021). This distinction is especially important in settings such as reading support, lexical ambiguity resolution, or context-dependent comprehension, where the usefulness of visual evidence depends not only on broad semantic relevance but on whether it clarifies the intended meaning of an expression in use. Visual signals that are broadly correlated with a concept may therefore be insufficient, or even misleading, when precise interpretation is required.

This distinction becomes particularly important in settings where multimodal information is intended to assist human comprehension or resolve ambiguity (Schneider et al., 2021; **Wang* et al.**, 2022). Here, grounding is not merely about recognizing shared content, but about identifying which aspects of visual evidence are relevant to a given expression under its contextual conditions. Prior research has made significant progress in scaling multimodal correspondence, yet the question of how to model context-dependent grounding—and how to construct resources that explicitly capture it—has received comparatively less focused attention.

Recent studies have begun to acknowledge this challenge by exploring more fine-grained alignments, structured annotations, and task-aware multimodal supervision (Plummer et al., 2015; Krishna et al., 2017; Hudson and Manning, 2019; Schwenk et al., 2022). These efforts suggest that effective grounding may require moving beyond generic image–text pairing toward data and learning strategies that encode when and why perceptual information is necessary for interpretation. Such a perspective motivates the investigation of multimodal resources and modeling approaches designed to capture the interaction between linguistic complexity, contextual usage, and visual support.

The work presented in Chapter 3 builds on this direction by examining how context-sensitive visual grounding can be operationalized through targeted dataset construction and evaluation, thereby complementing existing representation-focused approaches with a more usage-oriented view of multimodal learning.

2.2.2 Efficient Adaptation and Representation Integrity

The success of foundation models relies not only on large-scale pretraining but also on their ability to adapt efficiently to new domains and tasks. In practice, deploying such models often requires some form of specialization, whether through full fine-tuning, parameter-efficient adaptation, or prompt-based conditioning (Houlsby et al., 2019; Hu et al., 2022; X Liu et al., 2022; Poth et al., 2023; Shihab et al., 2026). A substantial body of research has therefore focused on reducing the computational cost of adaptation while maintaining competitive task performance. Techniques such as adapter modules, low-rank updates, sparse tuning, and other parameter-efficient fine-tuning strategies have demonstrated that effective customization can be achieved without modifying the entirety of a pretrained model (Houlsby et al., 2019; Hu et al., 2022; Sung et al., 2022; Jia et al., 2022; K Yao et al., 2026).

While these approaches have significantly improved the practicality of deploying foundation models, they typically evaluate success in terms of downstream accuracy

and efficiency gains (Poth et al., 2023; Han et al., 2024; K Yao et al., 2026). Less attention has been devoted to how adaptation affects the internal semantic structure inherited from pretraining. Because multimodal foundation models encode complex cross-modal relationships, even small updates can alter the geometry of the representation space, potentially weakening previously learned correspondences or reducing robustness under distribution shift (Luo et al., 2025; H Li et al., 2024; Shihab et al., 2026).

This tension highlights an important consideration: efficient adaptation is not solely an optimization problem, but also a question of representation integrity. A trustworthy model must retain the coherence of its pretrained knowledge while incorporating new information in a controlled manner. If adaptation disrupts the balance between modalities or introduces unintended drift, the resulting system may exhibit unstable behavior despite achieving strong task-specific results.

Recent work has begun to explore strategies that regularize adaptation through alignment-aware objectives, self-supervised constraints, uncertainty-aware adapters, or mechanisms that explicitly preserve relationships among representations (J Li et al., 2021; Bao et al., 2022; Sung et al., 2022; H Li et al., 2024; Shihab et al., 2026). These studies point toward a view of adaptation as a form of constrained transformation rather than unrestricted parameter updating, emphasizing continuity between pretrained and task-specific knowledge.

Chapter 4 contributes to this line of inquiry by investigating how lightweight adaptation can be guided to maintain both inter-modal alignment and intra-modal consistency, providing a pathway toward specialization that complements, rather than compromises, the representational foundations of multimodal models (**Wang et al.**, 2023).

2.2.3 Hallucination as an Inference-Time Phenomenon

As multimodal foundation models have become capable of open-ended generation, concerns about hallucination—producing content that is fluent yet unsupported by the input—have drawn increasing attention (Rohrbach et al., 2018; Y Li et al., 2023; Fu et al., 2025; Hanchao Liu et al., 2024; Bang et al., 2025). Prior research has investigated this issue from several perspectives, including improving training data quality, introducing factual supervision, and applying post-training alignment techniques (Sun et al., 2024; Gunjal et al., 2024; S Yin et al., 2024). These approaches have contributed valuable insights into how models acquire and express knowledge, and they have led to measurable improvements in reducing certain classes of errors.

However, hallucination is not solely a byproduct of training deficiencies. Generative models operate through autoregressive decoding processes in which multiple sources of information—learned priors, contextual prompts, and perceptual inputs—must be reconciled at each step of generation (Radford et al., 2019; Brown et al., 2020; XL Li et al., 2023). Even when representations are well aligned during training, the dynamics of decoding can amplify statistical biases or over-rely on language regularities, resulting in outputs that diverge from conditioning evidence (Chuang et al., 2024; Leng et al., 2024; Su et al., 2025). In this sense, hallucination can emerge from how models use their knowledge at inference time, rather than from what they have learned.

This observation has motivated a growing interest in examining generation as an interactive process that can be guided or regularized during inference. Instead of relying exclusively on additional parameter updates, recent studies explore decoding strategies, contrastive formulations, activation steering, and uncertainty-aware mechanisms that

adjust how competing signals are balanced when producing outputs (A Liu et al., 2021; XL Li et al., 2023; Chuang et al., 2024; Leng et al., 2024; Su et al., 2025; C Li et al., 2026). Such approaches suggest that faithfulness may be improved by directly shaping inference behavior, complementing training-based solutions with runtime control.

Viewing hallucination through this lens reframes it as a phenomenon arising at the intersection of representation and decision-making. Addressing it therefore requires not only better supervision, but also principled mechanisms for regulating how models translate internal representations into observable responses.

Chapter 5 builds on this perspective by investigating how inference-time contrastive strategies can mitigate hallucinations without retraining, illustrating how generation dynamics themselves can become a locus for improving trustworthiness (**Wang**, Pan, Ding, and Biemann, 2024).

2.2.4 Limitations of Safety Without Pragmatics

Ensuring that foundation models produce safe and socially acceptable outputs has become an important area of research, particularly as these systems are increasingly deployed in interactive settings (Bender et al., 2021; Weidinger et al., 2021). Existing approaches to safety often rely on filtering, controlled generation, reinforcement learning from human feedback, or post-processing strategies designed to reduce harmful or offensive language (Gehman et al., 2020; Schick et al., 2021; Ouyang et al., 2022; Rafailov et al., 2023). These methods have contributed substantially to mitigating explicit risks and have established practical mechanisms for moderating model behavior.

At the same time, many safety-oriented interventions operate primarily at the level of surface expression, treating undesirable outputs as patterns to be suppressed or replaced (Schick et al., 2021; Leong et al., 2023). While effective in reducing overt toxicity, such strategies may also alter the communicative properties of an utterance, including its emotional tone, stylistic nuance, or pragmatic intent. In conversational and user-facing applications, these aspects are not incidental; they form an integral part of meaning and influence how responses are interpreted by users.

This observation has led to increasing recognition that safety cannot be understood purely as the removal of problematic tokens or phrases. Instead, it involves managing a balance between acceptability and fidelity to the speaker's intent. Research has therefore begun to explore approaches that view detoxification as a constrained rewriting or transformation process, where the goal is to modify harmful elements while preserving semantic content and affective stance (Logacheva et al., 2022; Khondaker et al., 2024; Lee et al., 2024; Dementieva et al., 2025; **Wang**, Liu, et al., 2025).

Framing safety in this way highlights the need for models that can perform fine-grained, meaning-aware adjustments rather than global suppression. It also underscores the importance of evaluation protocols capable of assessing not only whether outputs are safe, but whether they remain contextually and pragmatically appropriate (Babakov et al., 2022; Guerini et al., 2013; Logacheva et al., 2022).

Chapter 6 contributes to this emerging perspective by examining sentiment-preserving rewriting as a mechanism for aligning safety with communicative intent, demonstrating how controllable generation can be formulated as a structured transformation rather than a purely restrictive process (**Wang**, Liu, et al., 2025).

2.2.5 Limitations of Interpretability Without Operability

Interpretability has become an important theme in the study of foundation models, motivated by the need to understand how large neural systems encode knowledge and arrive at their predictions (Belinkov, 2022; H Zhao et al., 2024; Rai et al., 2025). A wide range of approaches has been proposed, including probing methods that analyze linguistic or semantic information captured in hidden states, attribution techniques that estimate the contribution of input features, and visualization-based analyses that reveal patterns of attention or activation across layers (Geva et al., 2021; D Dai et al., 2022; Geva et al., 2022; Elhage et al., 2021). Collectively, these studies have provided valuable descriptive insights into the internal organization of large models.

Despite these advances, much of the interpretability literature remains primarily diagnostic in nature (Belinkov, 2022; Rai et al., 2025). That is, it seeks to explain model behavior after the fact, rather than to inform how models can be guided or modified in a principled way. Understanding which components encode certain types of information does not automatically indicate where interventions should be applied to achieve desired behavioral changes. As foundation models continue to scale, this gap between interpretation and operation becomes increasingly consequential for applications that require controllability, efficiency, or safety guarantees.

Recent work has begun to explore more intervention-oriented perspectives, investigating how insights from representation analysis can support targeted modifications, structured editing, or selective adaptation (N Zhang et al., 2024; Conmy et al., 2023; Ghandeharioun et al., 2024; Stolfo et al., 2023). These efforts suggest that interpretability can play a dual role: not only as a tool for explanation, but also as a basis for identifying actionable control points within deep architectures. Realizing this potential requires linking analytical signals to concrete mechanisms for steering model behavior.

This shift from descriptive to operational interpretability motivates approaches that integrate external, human-relevant signals with internal model structure, enabling interventions that are both interpretable and practically effective (Rayner, 1998; Hollenstein et al., 2020; Colman et al., 2022; **Wang**, Li, et al., 2024). Chapter 7 develops this idea by examining how cognitively inspired indicators can reveal functionally significant layers and guide selective interventions, transforming interpretability into a mechanism for controllable and efficient model steering (**Wang**, Pan, Ding, et al., 2025).

2.3 Summary

The preceding sections have first established the technical and methodological foundations needed to interpret the remainder of the dissertation. These foundations include transformer-based modeling, multimodal alignment and fusion, adaptation regimes, inference-time generation, decoding, and control, and selected human cognitive signals, particularly gaze during reading. The chapter has then reviewed several major research directions that contribute to the development of reliable foundation models, including multimodal grounding, efficient adaptation, hallucination mitigation, safety-oriented generation, and interpretability analysis. Each of these areas has produced important advances and addressed specific aspects of model behavior. At the same time, they have largely evolved along parallel trajectories, often focusing on localized improvements within particular stages of the modeling process.

Taken together, this body of work suggests that trustworthiness cannot be attributed to any single technique or component. Grounding methods strengthen the connection between perception and language, yet do not by themselves ensure stable adaptation. Efficient fine-tuning strategies enable practical deployment, but must be complemented by mechanisms that preserve semantic integrity. Efforts to reduce hallucination improve faithfulness at generation time, while safety-oriented approaches highlight the importance of maintaining pragmatic meaning. Interpretability studies, in turn, provide essential insights into model structure, but require translation into actionable forms of control.

These observations point toward the need for a more integrated perspective—one that treats trustworthiness as an emergent property of how data, representations, and inference mechanisms interact. Rather than addressing grounding, alignment, generation, or control in isolation, a bridged view seeks to understand how these elements can be coordinated to support dependable model behavior across diverse contexts.

This dissertation adopts such a perspective by investigating complementary solutions at multiple stages of the trustworthiness pipeline. The following chapters present a sequence of studies that operationalize this integration: constructing context-sensitive multimodal grounding resources, developing stable and efficient alignment strategies, regulating inference to enhance faithfulness, enabling meaning-aware safety interventions, and leveraging cognitively inspired signals for interpretable model steering. Through this progression, the dissertation aims to demonstrate how these interconnected efforts collectively contribute to the realization of trustworthy foundation models.

Use a picture. It's worth a thousand words.

— Arthur Brisbane (1911)

3

Contextualized Images for Complex Words to Improve Human Reading

Contents

Publication Note. This chapter is based on a first-authored publication by the dissertation author: Wang et al., "MOTIF: Contextualized Images for Complex Words to Improve Human Reading," published in the Proceedings of the 13th Language Resources and Evaluation Conference (LREC 2022). For integration into this dissertation, the material has been lightly revised for consistency in terminology, formatting, and cross-references. The core contributions and findings remain unchanged.

3.1 Abstract

MOTIF (MultimOdal ConTextualized Images For Language Learners) is a multimodal dataset that consists of 1125 comprehension texts retrieved from Wikipedia Simple Corpus. Allowing multimodal processing or enriching the context with multimodal information has proven imperative for many learning tasks, specifically for second language (L2) learning. In this respect, several traditional NLP approaches can assist L2 readers in text comprehension processes, such as simplifying text or giving dictionary descriptions for complex words. As nicely stated in the well-known proverb, sometimes "a picture is worth a thousand words" and an image can successfully complement the verbal message by enriching the representation, like in Pictionary books. This multimodal support can also assist on-the-fly text reading experience by providing a multimodal tool that chooses and displays the most relevant images for the complex words, given the text context. This study mainly focuses on one of the key components to achieving this goal; collecting a multimodal dataset enriched with complex word annotation and validated image match.

3.2 Introduction

Whether in human cognitive processes or computational systems, multimodal information is crucial for adequate concept formation, accordingly for language acquisition. Babies learn their native language by combining words with visual cues, e.g., the sound of the word "cat", an image of a cat, and a cat sound are all essential for the concept of *"cat"*. Over the past two decades, literature has provided convincing evidence on the facilitating role of cross-modal information in language acquisition (Ecalle et al., 2009; Dalton and Grisham, 2011; Hahn et al., 2014; Gerbier et al., 2018; Xie et al., 2019; Albahiri and Alhaj, 2020).

Although words are powerful symbolic representations, explaining the message (communicative intent) verbally yields unwieldy over specified sentences. Successful communication in daily communication settings usually involves linguistic information accompanied by other modalities like visual representations, gestures, or audio. The advantage of multimodal information holds for second language (L2) acquisition. Modern language learning applications or dictionaries like Babbel[1] or Duolingo[2] benefit from multimodality by using audio, visual illustrations, and video to enhance the L2 learning experience.

There are several approaches to assist non-native speakers in their reading activities. Through Lexical Simplification (LS), complex words can be replaced with simpler alternatives while preserving the meaning and syntactic function. It has been shown that LS leads to better text comprehension, improving text recall, especially for L2 learners at lower proficiency levels (Rets and Rogaten, 2021).

Instead of automatically simplifying the text, another approach would be to provide additional information about the complex word/phrase. This actively involves the reader by inviting her to process the supplementary/complementary information. Such a system can provide readers with dictionary definitions of the complex words in a more

1. https://www.babbel.com/
2. https://www.duolingo.com/

straightforward form. As we also address in this study, a more holistic approach can utilize multimodal information, e.g., an image, that depicts the information represented in the language modality. This would not only improve the understanding of the text but also facilitate the acquisition of new (complex) words by providing multimodal cues.

3.3 Application Scenario

Our ultimate goal is to provide language learners with a multimodal tool that chooses and displays the most relevant images for the complex words, given the context, to support their reading comprehension. To prevent any misunderstanding, a contextualized image should be chosen carefully to be in line with the information given in the rest of the sentence as much as possible. To achieve this, three central components should be addressed; (i) a multimodal dataset enriched with complex word annotation and contextualized images, (ii) complex word identification, and (iii) context-sensitive image retrieval. In this study, although we touch upon the last two items, we mainly focus on the dataset of comprehension texts for language learners enriched with images for the complex words. The texts provided in this dataset target English language learners below B1 proficiency level according to the Common European Framework of Reference for Languages (Council of Europe, 2001).

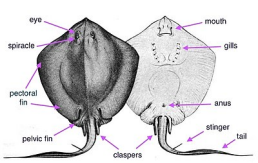
The following example illustrates what the model is expected to do. The text piece below[3] provides general information about stingray characteristics. Our focus is on their ability to camouflage in a sandy bottom, with the word “camouflage” as a complex word for our L2 readers.

> Stingrays use a wide range of feeding strategies. ... *Stingrays exhibit a wide range of colors and patterns on their dorsal surface to help them* ***camouflage*** *with the sandy bottom.* Some stingrays can even change color over the course of several days to adjust to new habitats. Since their mouths are on the side of their bodies, they catch their prey, then crush and eat with their powerful jaws.

Let’s assume that in our image pool, we have six images that a stingray is detected, as illustrated in Figure 3.1. While all the images are relevant at the surface level, they depict different concepts related to the animal "stingray," such as their different body parts or skin patterns (Figure 1a-d). Understanding the concept of the message is very crucial for providing a contextualized image that is more in line with what is explained in the sentence or the paragraph. For example, Figure 3.1e, which displays a harmless type of stingray, will be unfitting to explain the possible dangerous attacks. To improve the acquisition of the complex word “camouflage”, the system should be able to process the context and narrow it down the image selection to the image in Figure 3.1f. Providing any other image in such context may disrupt the reading fluency due to the conflict that it presents, or it may even yield misunderstandings of the text.

This study provides a semi-automatized dataset creation. First of all, existing established frameworks are used to detect complex words in the text. Second, state-of-the-art multimodal transformers are utilized to find a set of contextualized images for those words. Further, the match between the sentence, complex word, and image

3. retrieved from https://en.wikipedia.org/wiki/Stingray



(a) Drawing of a stingray anatomy

(b) Stingray belly side

(c) Spotted stingray

(d) Colorful spotted stingray

(e) A woman plays with stingrays

(f) Stingray lying on a sandy bottom

Figure 3.1: Supplementary image samples with simple descriptive captions for the word "Stingray". The images (without) caption are taken from https://en.wikipedia.org/wiki/Stingray

triple has been validated first by employing a crowdsourced platform and then by expert analyses as described in the upcoming sections. These semi-automatic detection methods incredibly reduce the costly and time-consuming manual annotation work.

This dataset has many other potential application areas in Language Education, Natural Language Processing, and Computer Vision, e.g., image-text alignment, sense-disambiguation. For example, when it is supported with psychological techniques such as eye tracking, the use of this corpus provides rich materials for researchers to investigate the mechanism and principles of multimodal learning in human beings and to develop computational models for them. It can also be used in building a more vivid learning context for children and L2 learners in some educational apps or courses.

3.3.1 Complex Word Identification

Detecting the complex words (CWI) in the texts is the first step towards providing L2 readers with assistance. CWI has received much attention in the past decade owing to SemEval 2016, 2018, and 2021 Shared Tasks that attract the attention of many NLP researchers to this domain (Paetzold and Specia, 2016a; Yimam et al., 2018; Shardlow et al., 2021). Complex words in texts can be identified by a wide variety of methods ranging from more traditional dictionary-based approaches to state-of-the-art deep learning techniques. Traditional approaches, which usually require domain knowledge and expert annotation, are still among the most common methods despite their costs. Using NLP approaches to detect complex words in a text helps minimize the manual work and mitigate the cost. Although there are end-to-end machine learning approaches for automatic complex word identification (CWI) (Paetzold and Specia, 2016b; Yimam

et al., 2018; Finnimore et al., 2019; Gooding and Kochmar, 2019) their success is still limited given the limited amount of data that have been trained on.

Although pre-trained language models can be used out-of-the-box on CWI tasks, fine-tuning on similar data is still very crucial to achieve better results on a specific task. Our multimodal dataset and semi-automatic data collection tools aim to close this gap. Therefore, our complex word identification will be more in line with the official proficiency standards or frameworks, such as the CEFR framework (Common European Framework of Reference for Languages) (Council of Europe, 2001). According to this framework, the proficiency levels range from A1 to C2. A1-level readers should understand very simple sentences and familiar words, while C1-level readers should comprehend a wide range of demanding, longer texts and recognize implicit meaning. Based on this coarse-grain classification, the L2 texts can be categorized into levels based on their vocabulary, such as (Uchida et al., 2018; Gooding and Kochmar, 2018). The details of the approaches will be elaborated on in the upcoming section.

3.3.2 Text-Image Retrieval

In the literature, text-image retrieval research has two directions; (i) text to image, i.e. image retrieval based on a textual query (ii) image to text, i.e. text retrieval based on an image query. However, in this work, we focus on textual queries and images as targets where the goal is to find the best matching images according to a sentence and a focus word within this sentence. Having an additional focus word in the query is an extension to common text-image retrieval and is described with more detail in Section 3.5. Nonetheless, in this work, we heavily rely on standard approaches described in the following text. Current state-of-the-art approaches for text-image retrieval are trained on multi-modal data comprising text-image pairs to compute the similarity between a text and an image. To find the best matching image, the models compute the similarity between the query and all images in the pool of images to be searched. Then the image with the maximum similarity to the query is selected as the best matching image. Current models are based on Transformer (Vaswani et al., 2017) architectures, and their inputs are textual tokens of a sentence and visual tokens of an image or, to be precise, their dense vector embeddings. Textual tokens embeddings are usually computed using pretrained transformer language models like BERT (Dementieva et al., 2025). Visual token embeddings are either regions-of-interest embeddings computed by pretrained object detection and classification models like Faster-R-CNN (Ren et al., 2015) or image-patch embeddings computed by a Vision Transformer (Dosovitskiy et al., 2021).

Despite having the same inputs, state-of-the-art models can be subdivided into two groups depending on how and when these two different modality representations are fused: early-fusion and late-fusion models. Early-fusion models like UNITER (Chen et al., 2020) or OSCAR (X Li et al., 2020) forward the textual and visual tokens through the same Transfomer-Encoder stacks, where a global text-image similarity score is computed via cross-modal self-attention. Despite their remarkable performance, early-fusion models are not applicable in real-time critical applications with large image pools because computing the similarity between a query and all images requires tremendous computational power. This is different for late-fusion models like TERAN (Messina et al., 2021) or VilBERT (Jiasen Lu et al., 2019), trained to compute joint representations of texts and images in a common vector space, typically by optimizing contrastive loss functions. To compute the representations, the models forward the input tokens through

two separated transformer-stacks – one for the textual and the other for the visual input. Then to compute a global similarity score, the outputs of the two transformer-stacks are fused in a cross-modal manner, individual on the model's implementation. This approach has the significant advantage that the image representations of all images in the pool to be searched can be precomputed so that only the query representation and the fusion of both have to be computed at inference time. In real-time critical applications with a large pool of images, this saves enormous amounts of time and computational power. While former late-fusion models generally perform worse than early-fusion models, the recent late-fusion model CLIP (Radford et al., 2021) achieves state-of-the-art performance. However, CLIP was trained on over 400M text-image pairs, which is significantly more training data than in all other mentioned models. Further, training CLIP requires massive GPU clusters due to the enormous batch sizes necessary during training. Fortunately, CLIP easily fits on a single consumer GPU during inference time. However, to collect the dataset presented in this chapter, we cannot use a traditional text-image retrieval approach as is since we extend the query by a contextualized focus word (aka complex word given the language proficiency level), which is part of the query sentence. When retrieving the best matching images, we additionally highlight the region in the image where the focus word is best represented according to the model – see Figure 3.4 for an example. This requirement originates from the language learner scenario, where we want to provide visual cues for complex words, which are the focus words in our dataset. More details of the context-depended image retrieval are described in Section 3.5.

3.4 Dataset Collection

A schematic overview of the data collection pipeline is depicted in Figure 3.2. More details on single steps are described in the respective sections.

First, the sentences from the Simple Wiki Dataset (c.f. Section 3.4.1) are tokenized using the NLTK tool[4]. Then, we conduct lemmatization in the pre-processing step to convert the inflected forms of each word. After that, each token is tagged with respect to its complexity (c.f. Section 3.4.3) and depictability (c.f. Section 3.4.4). If a token is both complex and depictable, it is marked as a focus word. The sentences containing less than three focus words are discarded to ensure a level of complexity. The result is a set of samples, where each sample consists of a context sentence and a focus word (a word that will be supported by a contextualized image). Next step is the context-dependent image retrieval. In this stage, the top-5 matching images from MS COCO (Lin et al., 2014) are retrieved, and the focus word region is highlighted in the image with a boundary box (c.f. Section 3.5).

Since the final dataset should only contain the best image that perfectly matches the context and focus word, additional filtering stages are employed. First image filtering stage has been conducted automatically by using a state-of-the-art multimodal transformer model. A pretrained and publicly available CLIP model (Radford et al., 2021) is used to compute the cosine similarity of each context sentence and the retrieved top-5 images. Samples are discarded, where not all images have a similarity score of at least 0.225. This was inspired by the LAION-400M dataset (Schuhmann et al., 2021)

4. https://www.nltk.org/

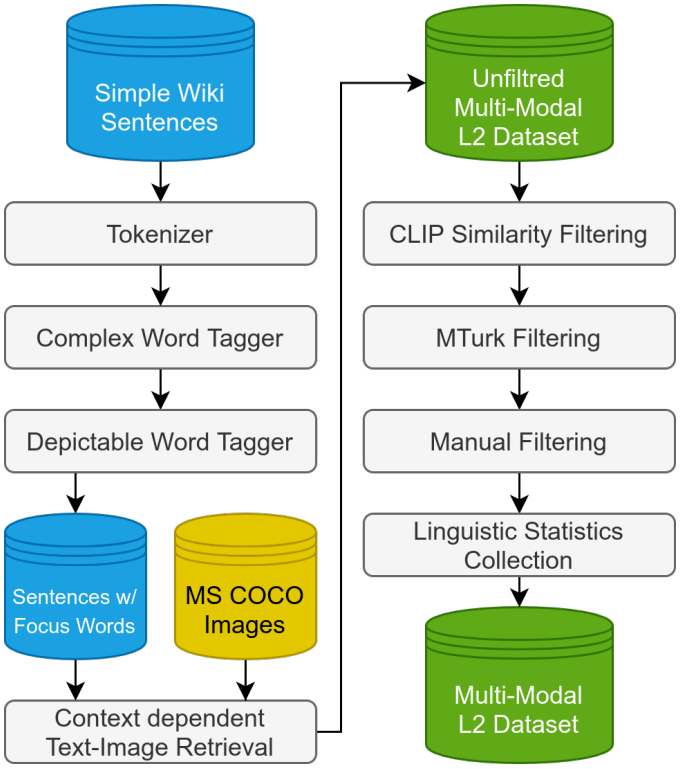


Figure 3.2: Schematic overview of the dataset collection pipeline.

with slightly looser similarity requirements. To further increase the quality of the text image pairs and to make sure that the highlighted image region matches the focus word, we conducted crowdsourcing experiments on Amazon MTurk[5] where human workers are asked to rate how well the respective images match the corresponding focus and context (c.f. Section 3.6).

After that, in the manual next filtering stage, the authors hand-selected the best matching image from every sample and dropped samples where none of the images represented the context and focus word well enough. With this filtering step, we ensure that only the highest quality samples are included in the final dataset. Since it might be useful in various use case scenarios and downstream research to have linguistic statistics like the number of tokens, POS tags, or named entities in a sample, we collect those using a spaCy[6] powered pipeline and released with the rest of the dataset.

3.4.1 L2 Learner Reading Material

It is essential to gather appropriate texts for L2 learners that target L2 word acquisition. In order to include proper texts in this dataset, we define several criteria that the text should meet. Firstly, the topic of a text should be open domain (not created based on pre-defined templates), such as the paragraph about stingrays extracted from Wikipedia (Section 3.3). Unlike widely used multimodal datasets, whose text pieces are merely daily contents, our text pieces cover a wider variety of topics, such as art, culture, geography, nature, science, technology, etc. Furthermore, the text structure should also display complexity. Specifically, the average tokens per paragraph (the text length)

5. https://www.mturk.com
6. https://spacy.io/

should be more than the average tokens in typically used captions. Meanwhile, the comprehension level for a given text should be aligned with the respective reading level for L2 learners. The last criterion is to have *named entities* in the text. Constrained by their annotation method, which first provides an image then asks annotators to write sentences, caption-based datasets exclude *name entities*. However, *named entities* commonly appear in reading materials, such as biography, geography, etc., and thus they are an essential part of reading materials. For this reason, unlike the existing multimodal datasets, our L2 text should involve *name entities*. Motivated by these criteria, we have pre-processed text from five different sources, which is elaborated in the Future Direction section. In this study, we choose the Wikipedia Simple Corpus as the textual part for the following procedures considering both its scale and related characters that match our hypotheses mentioned above.

Wikipedia Simple Corpus is the dataset collected from Simple English Wikipedia [7]. There, editors use simple English words and grammar but contain the same entries and content resulting 201.531 articles, which are suitable for children and L2 learners as compared to Normal English Wikipedia [8]. We have utilized the sentences from the raw Wikipedia Simple Corpus (Benzahra and Yvon, 2019), which is a single file containing 505.974 paragraphs where several consistent paragraphs belong to an identical article. After processing the raw data, 59.769 unique articles are kept in this phase. For each article, the average number of paragraphs is *eight*, while for each paragraph, the average number of tokens is *eighteen*. Besides, we compute the average score for the Flesch–Kincaid readability tests (Kincaid et al., 1980) to assess the readability grade level of this corpus. In the Flesch reading ease test, higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. The average Flesch–Kincaid readability for Wikipedia Simple Corpus is 64.2, which means the reading materials are suitable for 13-15 US school students in grades 8-9.

3.4.2 Supplementary Images

The source of the images (Lin et al., 2014) used for this dataset is retrieved from the 2017 version of MS COCO, a popular dataset for various computer vision tasks on natural, non-iconic images. It comprises about 123K carefully selected, annotated, and captioned images from Flickr[9]. We chose this dataset because the images show a wide range of different objects and scenes and are further under the Creative Commons License, which allows non-commercial use and distribution.

3.4.3 Complex Word Tagger

CWI is developed as a fundamental prerequisite for lexical simplification (LS). However, as described in section 3.3.1, providing a simplified version is a different task than providing additional information to enhance the acquisition of the complex word. Therefore, the annotated data in the existing datasets created for LS commonly tend to be rare words or long phrases. Due to this inconsistency, in our current study, we prefer to use an established CEFR framework designed for language proficiency to annotate the complex words in our datasets instead of deep learning approaches.

7. https://simple.wikipedia.org/
8. https://www.wikipedia.org/
9. https://flickr.com

The CEFR framework describes the skills learners should develop at each of the six proficiency levels of the scale. But, it doesn't provide a word list directly with corresponding levels. To overcome this limitation, we obtain a word list with related CEFR level labels by dealing with a word frequencies list released by EFLLex (D.ürlich and François, 2018) in the lexical learning domain. In the EFLLex, word frequencies are collected from materials designed for English (as L2) learners, which contain 15, 282 words. In this frequency list, same words with different *part-of-speech* tags are listed separately. For simplification, we combine the frequencies of a word with different POS tags. Then, this list has 9, 396 unique words with frequencies from A1 to C2. To transform the word's frequency to the corresponding CEFR level, we adopt the strategy that the level containing the most significant frequency is set as the proficiency level for a particular word. B1 level, which entails that the learned language can be used freely in daily study, and work scenarios, is chosen as a threshold. Thus, we label a word as a complex word if its CEFR label is above B1 level. As a result, 1690 words are labeled as easy words, while 7706 words are marked as complex, in proportion to 82% of the words in the list.

3.4.4 Depictability Tagger

Unlike previous work in constructing multimodal datasets, we also take the depictability of words into account. The word *dog* is more depictable than the word *beautiful*. Learners can comprehend complex words better with visual clues if visually depictable words exist in the context. Besides, we seek to conduct a context-dependent text-image retrieval task in this study. Our model can obtain the most relevant images paired for given sentences only if visually depicted elements exist in sentences. To this end, after tagging the complex words, tokens are labeled concerning their depictability by the depictable word tagger (a binary classification; *yes* or *no*).

(Brysbaert et al., 2014), in their psycho-linguistically motivated research, ask native speakers to label 40K words using the Amazon Mechanical Turk platform. In their work, a 5-point rating scale is used to rate a word as abstract or concrete, where 1 means abstract most, whereas 5 means concrete most. The concreteness parameter corresponds to the concept of *depictability* in our research. Finally, 2.3M ratings were collected, and they released the average rating score for each token. Using this rating list for 40K English lemmas[10], we compute the depictability score for each token with the min-max normalization equation below.

$$d_{score} = \frac{r_{token} - r_{min}}{r_{max} - r_{min}}$$

where r_{token} is the average rating point under the 5-point rating scale for a given token, and r_{min} and r_{max} are minimum and maximum rating points in the list respectively. More specifically, Figure 3.3 shows the depictability score distribution for words after the normalization.

Object and attribute labels of MS COCO dataset obtained from the Faster R-CNN model (Ren et al., 2015) are another source to label a word as depictable, because these labels are mostly noun words whose objects are appeared in images. There are 1625

10. http://crr.ugent.be/archives/1330

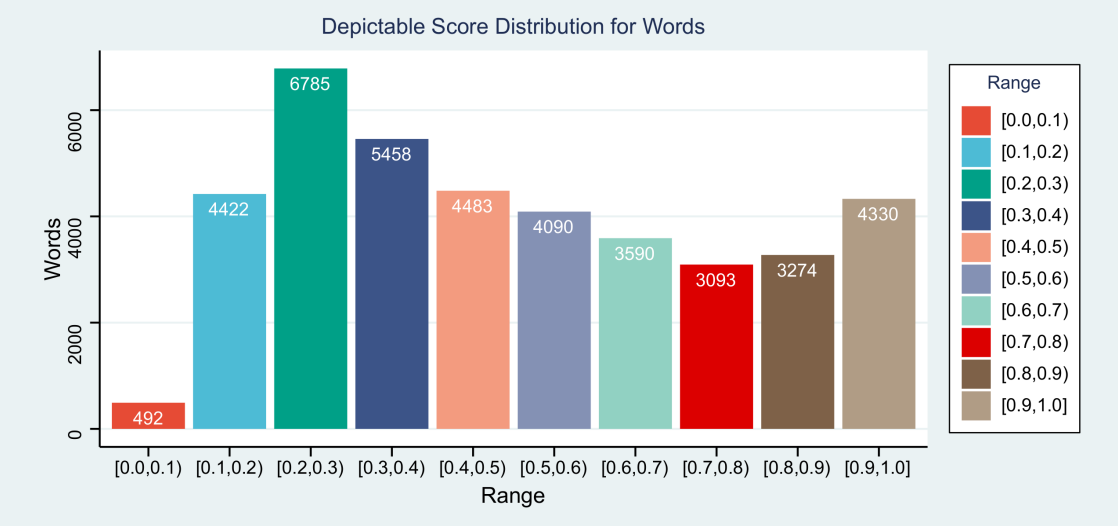


Figure 3.3: A bar chart of depictable score distribution for 40K English lemmas. We group tokens conditioned on the depictable scores into ten ranges, from $[0.0, 0.1]$ to $[0.9, 1.0]$. This figure shows that around 10.84% of tokens whose depictable scores are over 0.9.

tokens in the object label list of the MS COCO dataset. In the end, for each word in the sentences, we apply the equation below to get respective depictable labels.

$$label_{depictable} = \begin{cases} 1 & \text{if } d_{score} \geqslant 0.9 \text{ or } \in l_{object} \\ 0 & \text{if } d_{score} < 0.9 \text{ or is } OOV \end{cases}$$

where d_{score} is the depictability score for a given token, l_{object} is the list of object labels, OOV is a token out of these two lists, and 0.9 is the threshold score to decide a word as depictable or not.

By utilizing both complex and depictable word taggers, a word in the textual modality is set as a focus word if its hard label and depictable labels are both positive. At the same time, in the visual modality, a focus word is the object label detected by an object detection model.

3.5 Context-dependent Image Retrieval

As an extension to the common text-image retrieval task introduced in Section 3.3.2, where the best matching images for a textual query consisting of a sentence or a word must be found, we introduce context-dependent text-image retrieval in previous work (Schneider, 2021). The difference is that the query is a pair that comprises a sentence, referred to as context, and a focus word contained in the sentence. Further, the goal is to retrieve the best matching images regarding the context with particular attention to the focus word within the context and find the image region where the focus word is represented best (c.f. Figure 3.4).

To accomplish this goal, we use a pretrained TERAN model for standard text-image retrieval and apply a re-ranking stage to attend to the focus word specially and to find the region where the focus word is represented best. TERAN is a late-fusion model that computes the global similarity between an image and a textual query – in our case called context – by aggregating a fine-grained word-region-alignment (WRA) matrix

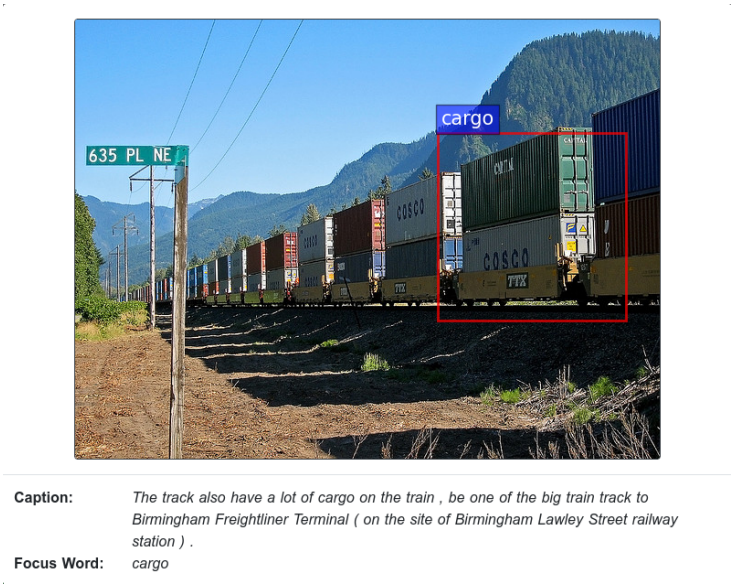


Figure 3.4: An example of a context-dependent image retrieval query with the best matching image where the focus word region is highlighted. The query comprises a context sentence (referred to as "Caption" in this figure) and a focus word with the context sentence.

A. The cells of **A**, are the cosine-similarities of the visual regions of the image I and textual tokens of the context sentence C are defined as

$$\mathbf{A}_{i,j} = \frac{\mathbf{v}_i^T \mathbf{t}_j}{|\mathbf{v}_i||\mathbf{t}_j|}$$

where $\mathbf{v}_i \in I$ and $\mathbf{t}_j \in C$.

The global similarity, i.e., the "context-score" $s_{context}$, of an image and a context sentence is defined as

$$s_I^{(c)} = \sum_{j \in |C|} \max_{i \in |I|} \mathbf{A}_{ij}$$

To specially attend to the focus word, we first compute a "focus-score" s_{focus} based on the WRA matrix.

$$s_{focus} = \frac{1}{N * (f_e - f_s + 1)} \sum_{i=0}^{N} \sum_{j=f_s}^{f_e} \mathbf{A}_{ij}$$

where N is the number of regions per image; f_s and f_e are the starting and ending indices of the focus in the context, respectively; and **A** is the WRA matrix of an image I and the context C.

After that, we first normalize and then combine the global similarity (interpreted as the "context-score") with the "focus-score" with a weighted average to obtain the image score $s_{combined}$ for the context-dependent text-image retrieval.

$$s_{combined} = \alpha \cdot s'_{context} + (1 - \alpha) \cdot s'_{focus}$$

where $\alpha \in [0, 1]$ is the weight for the weighted average; $s'_{context}$ and s'_{focus} are the normalized "context-score" and the "focus-score", respectively. For the dataset presented in this chapter, we set $\alpha = 0.9$.

The image with the highest score is the best matching image according to the context and focus word. To highlight the region where the focus word is represented best, we select the region with the maximum "focus-score".

3.6 Crowdsourcing Experiments

Since the context-dependent image retrieval stage does not always retrieve images representing the context and the focus flawlessly, a crowdsourcing experiment was conducted to filter out these samples with the aim of increasing the dataset quality. In this experiment held on Amazon MTurk, workers were given the task to rate how well the context and the focus word are represented in the corresponding image and highlighted image region, respectively, on a 5-star scale. Because the default questionnaires available on MTurk do not efficiently support this task, a tool including a custom web application was developed. Using the "External Question" of MTurk, access to the web application was provided within the MTurk Marketplace environment.

The data for the study comprised 3125 samples, each consisting of a context sentence, a focus word, five images, and two 5-star scales for the context and focus word, respectively, for each of the images. Like in an image slideshow, the workers can switch between the images so that only one image and the corresponding 5-star scales are shown at a time. An example of the application UI as it is presented to the workers is shown in Figure 3.5.

Figure 3.5: An example of the crowdsourcing experiment UI for MTurk workers.

Using our tool, the samples were published on the MTurk marketplace as HITs. To ensure high-quality results, which are not biased by the opinion of single users,

we require three assignments of three different workers per HIT. Further, to accept a HIT, a worker needs at least 1000 approved assignments and an approval rate of 90%. Considering ethical fairness, we set the reward per assignment to 0.2€. With this, an estimated duration of one minute per assignment results in an hourly salary of 12€.

After receiving all assignments for all HITs, the results were filtered as described in the following. First, a score for each image j of the top-5 images per sample i is computed based on the focus rating f_w and the context rating c_w of the three workers

$$\textbf{score}_i^j = \sum_{w=1}^{3} \max(f_w - 4.0 + c_w - 4.0, 1.0)$$

Then, images are dropped if their score is below or equal to a threshold $T = 2.0$ In other words, an image j of a sample i is kept if at least two workers rated with at least 4.0 stars that the focus and the context is represented well in the respective image. After that, only samples with at least one well-rated image are kept for the manual selection stage.

The authors further filtered down the selected sentence-focus word-image triplets in the last manual stage to ensure the quality of the dataset (*manual expert annotation*). In this stage we discarded 531 samples.

3.7 Dataset Structure and Statistics

The original Wikipedia Simple Corpus contains $506K$ sentences. To improve the accuracy of the following retrieval task, we set the threshold of the focus word numbers in each sentence as three. After the complex word tagger and depictable word tagger mentioned in Section 3.4, $31K$ samples with at least three focus words are kept to be used to conduct the context-dependent image retrieval task discussed in Section 3.5. At last, after crowdsourcing experiments and manual expert annotation steps, we got 1125 text samples paired with the best matching images for the complex words.

Each sample in the final dataset is contains:

- a sentence referred to as context
- a word within the context referred to as focus word that is both complex and depictable
- an image that globally represents the context and represents the focus word in a highlighted bounding box
- linguistic statistics about the context such as the number of tokens and their respective POS tags

To sum up, there are 1125 samples, where we have 695 unique context sentences and 277 unique focus words. The average tokens per paragraph are 28, while the minimum and maximum tokens per paragraph are 8 and 94, respectively. Meanwhile, we compute the average ratio of tokens associated with name entities vs. all tokens as 3.84%. Besides, the average Flesch–Kincaid readability score is 72.39, interpreted as students' ability in $7th$ Grade in US school. Last, numbers of tokens with different POS tags, as shown in Figure 3.6. Two visual samples of the final MOTIF dataset are shown in Figure 3.7. The dataset is available on an open-access repository via the link.

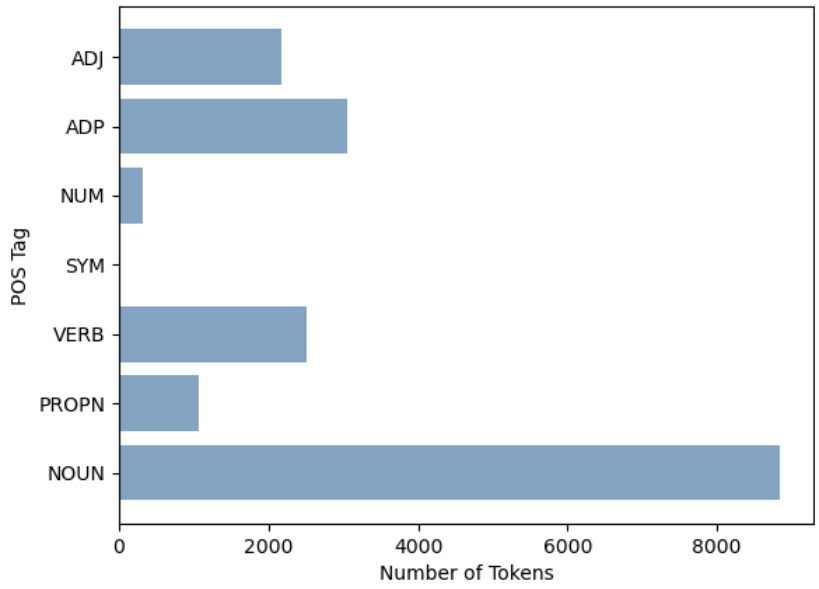


Figure 3.6: Number of tokens in MOTIF with different POS Tags. NOUN, PRON, VERB, SYM, NUM, ADP, and ADJ mean noun, pronoun, verb, symbol, numeral, adposition, and adjective words respectively.

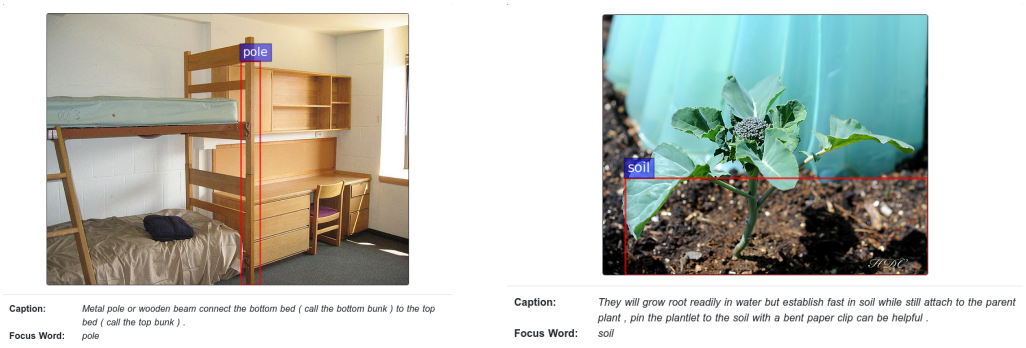


Figure 3.7: Visual examples included in the MOTIF dataset.

3.8 Future Directions

There are various ways to extend the dataset in future work. We plan to use additional text-only L2 learner reading material and forward it through the dataset collection pipeline (c.f. Figure 3.2) to increase the number of samples in the final dataset. Possible resources for this are, e.g., Wikipedia Normal (Benzahra and Yvon, 2019), InScript (Modi et al., 2016), Weebit (Chen and Meurers, 2016), or OneStopEnglish (Vajjala and Lučić, 2018). These datasets vary in size, topics, and length of the sentences but are all specially designed to be understood by 6th to 12th-grade students from US schools.

Further, several components in the dataset collection pipeline can be enhanced to improve the efficiency of the dataset collection and the quality of the final dataset. To begin with, since CEFR word-complexity classification is conducted at the single

token level, the use of a simple tokenizer was sufficient. However, this, unfortunately, rips apart multi-word expressions (MWEs) like compound nouns. The complex word tagger can be improved by adapting state-of-the-art complex word identification (CWI) approaches and resources (Kochmar et al., 2020) which pay special attention to multi-word expressions (MWE). Moreover, this dataset can be used to fine-tune SOTA CWI models to improve automatic complex word detection.

Further, the employed object and attribute vocabulary, which comprises about 1600 different terms, can be significantly extended by the vocabulary of the Visual Genome dataset, which contains about 75K unique object types and about 40K attribute types. By improving the pipeline as described, the quality of the output of the context-dependent image-retrieval stage will automatically increase. However, the stage itself can be further improved by various methods briefly summarized in the following.

The currently employed context-dependent image-retrieval model is a TERAN model, where the visual inputs are region-of-interest (ROI) feature vectors computed by a pre-trained Faster-R-CNN model. The advantage of this approach is that we can compute the focus score (c.f. Equation 3.5) of a focus word and an image (region) from the WRA matrix, which holds fine-grained cosine similarities between words and image regions. However, the bounding box of the image region often does not perfectly fit the underlying object representing the focus word.

To resolve these issues, we plan to leverage a pre-trained CLIP model in the context-dependent image-retrieval stage. We will utilize class activation mapping techniques introduced (B Zhou et al., 2016; Selvaraju et al., 2019) to compute the focus score and more accurate bounding boxes.

3.9 Conclusion

In this study, we present a semi-automatized pipeline to create a high-quality multi-modal dataset containing text pieces for L2 speakers, annotated complex words, and contextualized images that ease the comprehension of the complex word given the context. Our pipeline starts with selecting L2 text, conducting text analysis (number of *named entities*, readability scores etc). It further detects complex words using well-established CEFR levels and employs SOTA NLP approaches for finding contextualized images. However, these automated processes are followed by a careful validation method using the Amazon MTURK crowdsourcing platform and expert analysis. The resulting dataset consists of 1125 text samples annotated with complex words and context-dependent images for these words.

This multimodal support approach and the dataset are not only for the L2 domain, but they can also be used in developing assistive systems for people with low literacy and reading difficulties. Further, these enriched annotations can be instrumental in fine-tuning or testing automatic CWI and contextualized image-retrieval models.

Seeing comes before words.

— John Berger (1972)

4

Using Dual Constraint Contrastive Learning for Cross-modal Retrieval

Contents

Publication Note. This chapter is based on a first-authored publication by the dissertation author: Wang et al., "Using Self-Supervised Dual Constraint Contrastive Learning for Cross-Modal Retrieval," published in the Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023). For integration into this dissertation, the material has been lightly revised for consistency in terminology, formatting, and cross-references. The core contributions and findings remain unchanged.

4.1 Abstract

In this work, we present an unsupervised dual constraint contrastive method for efficiently fine-tuning the vision-language pre-trained (VLP) models that have achieved great success on various cross-modal tasks, since full fine-tune these pre-trained models is computationally expensive and tend to result in catastrophic forgetting restricted by the size and quality of labeled datasets. Our approach freezes the pre-trained VLP models as the fundamental, generalized, and transferable multimodal representation and incorporates lightweight parameters to learn domain and task-specific features without labeled data. We demonstrated that our unsupervised dual contrastive model performs better than previous fine-tuning methods on MS COCO and Flickr 30K datasets on the cross-modal retrieval task, with an even more pronounced improvement in zero-shot performance. Furthermore, experiments on the MOTIF dataset prove that our unsupervised approach remains effective when trained on a small, out-of-domain dataset without overfitting. As a plug-and-play method, our proposed method is agnostic to the underlying models and can be easily integrated with different VLP models, allowing for the potential incorporation of future advancements in VLP models.

4.2 Introduction

With the rapid growth of computational power and extensive large-scale data, increasingly advanced foundation models have been proposed in both the language domain (Devlin et al., 2019; Y Liu et al., 2019; Touvron, Lavril, et al., 2023) and the vision domain (Dosovitskiy et al., 2021; Dehghani et al., 2023). By leveraging these breakthroughs as the backbone, vision-language pre-trained (VLP) models have made significant strides in a range of cross-modal tasks (J Li et al., 2021; J Yu et al., 2022; Bao et al., 2022; W Wang et al., 2023), demonstrating that multimodal representations derived from pre-trained models possess exceptional generalization and transfer capabilities.

In line with the successes of VLP models, recent works (T Yang et al., 2023; Sung et al., 2022; Diao et al., 2023) have adopted the "pre-training and fine-tuning" paradigm for downstream cross-modal tasks and out-of-domain scenarios. As shown in Figure 4.1, there are two prevalent fine-tuning strategies. The first, full fine-tuning, involves fine-tuning all parameters, but it carries two notable drawbacks: computational efficiency and catastrophic forgetting (J Li et al., 2022). Given the substantial number of parameters in VLP models, considerable memory is required to store these parameters, not to mention train the entire model. For example, the CLIP model (Radford et al., 2021) utilized 592 V100 GPUs over a span of 18 days. Furthermore, in the absence of high-quality labeled datasets, fully fine-tuning VLP models often results in catastrophic forgetting (J Li et al., 2022), where the previously learned generalized and transferable multimodal representations from VLP models degrade. The second method, frozen and fine-tuning, offers greater flexibility by freezing VLP model parameters while adding blocks on top to learn out-of-domain and task-specific representations. To achieve state-of-the-art performance on benchmark datasets, these extra blocks tend to be sophisticated and task-specific tricks have been proposed. For instance, in the cross-modal retrieval task, state-of-the-art approaches heavily rely on region feature extraction (Girshick, 2015), cross-modal fusion (Jiasen Lu et al., 2019), and hard negative sampling (Faghri et al., 2018) during fine-tuning. (Rao et al., 2022) reveals that while these techniques are

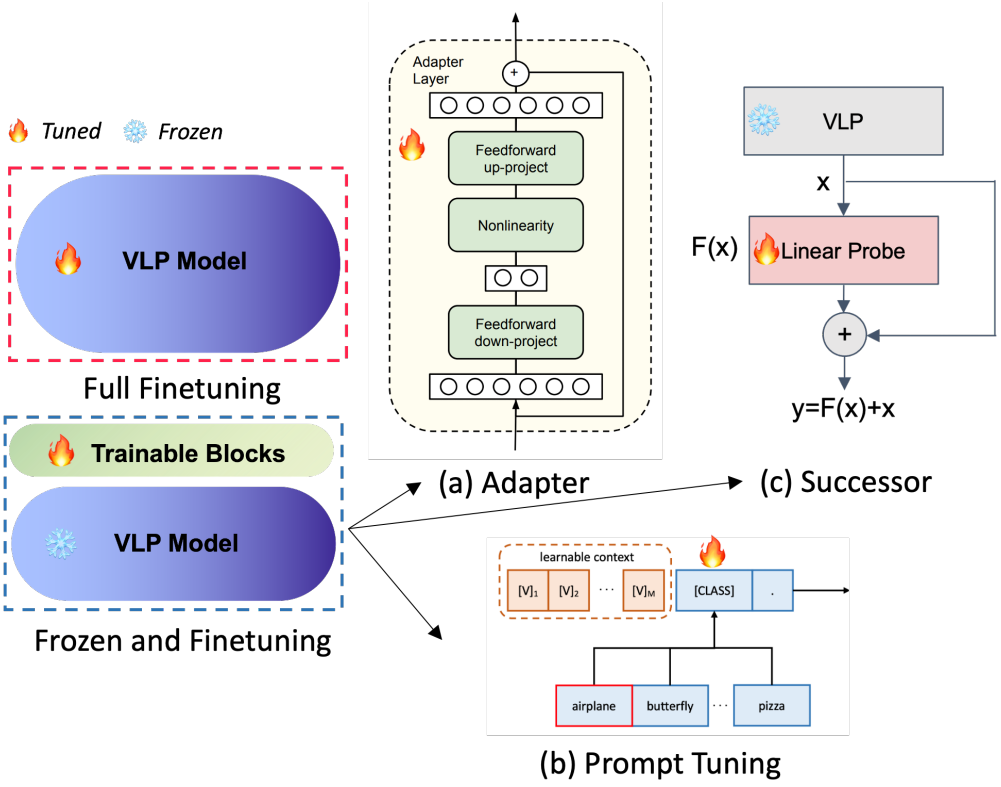


Figure 4.1: Pre-training and fine-tuning paradigm: full fine-tuning and frozen and fine-tuning.

crucial for improving performance on benchmark datasets, they come at the cost of increased training time, reduced efficiency, and diminished transferability and utility when applied to different domains.

To address the challenges mentioned earlier, following frozen and fine-tuning, Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019) has recently gained popularity and attracted significant interest. The core idea of PEFT is to utilize a smaller set of parameters for fine-tuning while retaining the capabilities of pre-trained foundational models to improve transferability and adaptability. Among these PEFT approaches, adapters (Houlsby et al., 2019) add and update new parameters at the model level, while prompt tuning methods (K Zhou et al., 2022) incorporate and train parameters at the input level. Although these techniques have proven effective, they remain inadequate when VLP models are not available for adapter injection, and considerable effort is needed to identify the best prompt templates. In most cases, paired multimodal datasets are not readily accessible. We propose that an optimal solution would involve adding additional parameters at the output level and training the extra layers in an unsupervised manner, without relying on any tailored techniques.

This study introduces an unsupervised dual constraint contrastive learning for cross-modal retrieval task (SUCCESSOR), inheriting the ability of VLP models. In various cross-modal tasks, dual attributes exist (Qin, 2020). For example, if the primary task in cross-modal retrieval is text retrieval, the dual task would be image retrieval. We construct a dual constraint contrast in the primary modality by back-retrieving negative samples from the dual modality and vice versa, aiming to enhance the alignment of multimodal representations within both intra- and inter-modalities. Specifically, beginning with the primary modality (e.g., vision), we perform forward retrieval (text

retrieval) to obtain negative samples from the dual modality (language). We then use these retrieved negative samples to conduct back-retrieval (image retrieval), acquiring candidates in the primary modality. This process allows us to compare the semantic distances between the candidate and original query in the prime modality and vice versa, thereby increasing the alignment and coherence of the multimodal representations.

In terms of our model, we freeze the VLP models to serve as the foundational generalized multimodal representations and add two linear probe layers on top to learn out-of-domain and task-specific representations involving super lightweight parameters for fine-tuning. A skip shortcut is introduced to connect the in-domain representations with the final output of the linear probes, facilitating rapid tuning and model convergence. Our experiments demonstrate that the unsupervised SUCCESSOR model, without relying on region feature extraction or any hard negative sampling techniques, can compete with fine-tuning methods on benchmark datasets such as MS COCO (Lin et al., 2014) and Flickr 30K (Plummer et al., 2015). Surprisingly, we discovered that random in-batch negative sampling offers a diverse choice of negative samples, enabling the model to learn fine-grained multimodal semantics, rectify errors from VLP models, and ultimately enhance cross-modal retrieval performance.

Owing to the simplicity of our proposed method, fine-tuning can be completed within hours on an A6000 GPU (48 GB) and can function as a plug-and-play approach, easily integrating with various VLP models without the need for labeled paired data. This adaptability allows for the potential incorporation of future advancements in VLP models. To summarize, our contributions are as follows:

- We introduce a new PEFT approach—an unsupervised dual constraint contrastive method—by adding lightweight, learnable parameters at the output layers. Our method is cost-effective, requiring only a single GPU and a few hours for fine-tuning without the need for labeled datasets, functioning as a plug-and-play solution.

- Our unsupervised method achieves comparable or superior performance to previously fine-tuned state-of-the-art methods on standard benchmark datasets, such as MS COCO and Flickr 30K, without relying on region feature extraction, complex cross-attention fusion, or hard negative sampling strategies.

- By freezing the parameters of VLP models and introducing a skip shortcut, our method yields fast convergence while preserving the generalization and transferability of VLP models. Zero-shot experiments demonstrate that SUCCESSOR further improves cross-modal performance accuracy compared to the VLP backbone, showcasing that SUCCESSOR inherits VLP capabilities.

- A domain adaptation experiment on the education-oriented, small dataset MOTIF (**Wang* et al.**, 2022) reveals that SUCCESSOR performs effectively in domain adaptation without overfitting.

4.3 Related Work

Vision-language pre-training: We are witnessing an era in which advanced foundational models rapidly evolve in visual and language modalities (Dosovitskiy et al., 2021;

Dehghani et al., 2023; Devlin et al., 2019; Y Liu et al., 2019; Brown et al., 2020; Touvron, Lavril, et al., 2023). In line with the advancements in unimodal foundational models, VLP models have garnered significant research interest. Early models such as ViLBERT (Jiasen Lu et al., 2019) employed a dual encoder and cross-attention to learn multimodal representations, while UNITER (Chen et al., 2020) and OSCAR (X Li et al., 2020) utilized a fusion encoder with self-attention to learn multimodal alignment. ViLT (Kim et al., 2021) argued that visual patches from vision transformers are more efficient and enable end-to-end model training. More recently, CLIP (Radford et al., 2021) adopted large-scale multimodal data from the internet and employed a contrastive method for training, resulting in more powerful multimodal representations and impressive zero-shot performance. Meanwhile, ALBEF (J Li et al., 2021) demonstrates that image-text contrastive, masked language modeling, and image-text-matching tasks are more efficient than other pre-training tasks. To enhance multimodal generation capabilities, models like BLIP (J Li et al., 2022), Flamingo (Alayrac et al., 2022), and CoCa (J Yu et al., 2022) have been proposed, enabling VLP models to handle both multimodal understanding and generation tasks. Most recently, the VLMo model (Bao et al., 2022) introduced multiway transformers, unifying the dual encoder and fusion encoder approaches. Building on VLMo, the BEiT-3 model (W Wang et al., 2023) has achieved new state-of-the-art results on cross-modal learning benchmark tasks and even single-modality tasks. We opted for the CLIP model as our VLP model due to its demonstrated efficiency in generalized multimodal feature extraction, moderate parameter size, and the fact that it does not necessitate a pre-trained Fast-RCNN model (Girshick, 2015). Given that our proposed method is a plug-and-play solution, we believe it can be easily applied to other VLP models and even future advancements in the field of VLP.

Parameter-efficient fine-tuning: There are two widely-used fine-tuning approaches: full fine-tuning and frozen fine-tuning. Full fine-tuning presents two drawbacks: computational efficiency and catastrophic forgetting (J Li et al., 2022). Given the large number of parameters in VLP models, training the entire model becomes less feasible. Moreover, without high-quality labeled datasets, fully fine-tuning VLP models can lead to catastrophic forgetting (J Li et al., 2022), where the previously learned generalized and transferable multimodal representations from VLP models deteriorate. In contrast, frozen fine-tuning offers more flexibility and strikes a balance between accuracy and the number of trained task-specific parameters. Recently, parameter-efficient fine-tuning (PEFT) (Houlsby et al., 2019) has gained popularity following the frozen fine-tuning fashion. Among PEFT approaches, adapters (Houlsby et al., 2019) introduce and update new parameters at the model level, while prompt tuning methods (K Zhou et al., 2022; Jia et al., 2022) incorporate and train parameters at the input level. Although these techniques have proven effective, they remain inadequate when VLP model training codes are unavailable for adapter injection, and significant effort is required to identify the best prompt templates. We believe that an optimal method involves adding additional parameters at the output level, using VLP models as the fundamental multimodal representation and fine-tuning the extra parameters to learn out-of-domain and task-specific representations.

Cross-modal retrieval: Cross-modal retrieval, such as image-text retrieval (Rao et al., 2022; Diao et al., 2023) requires accurate alignment and understanding of information from different modalities, making it an ideal task to evaluate the performance of our unsupervised dual constraint contrast method. Past research has focused on various ways to improve results on benchmark datasets like MS COCO and Flickr 30K. Although

multiple state-of-the-art methods have been proposed to achieve SOTA results on these datasets, they can be categorized into three main directions. First, for instance, region features (Girshick, 2015) are crucial for improving accuracy in the visual modality (Rao et al., 2022), while BERT (Devlin et al., 2019) features outperform RNN features. However, obtaining region features is less efficient and requires pre-training object detection modules (Girshick, 2015). Visual patch projection (Dosovitskiy et al., 2021) is more efficient as it allows for end-to-end model training. Second, fusion encoder (Chen et al., 2020) use self-attention to learn the interaction between modalities, while dual encoders (Jiasen Lu et al., 2019) employ cross-attention to interact with different modalities. Lastly, techniques like in-batch hard negative mining (Faghri et al., 2018) have proven effective in increasing the relevance score between paired data while decreasing the score for non-paired data. (Rao et al., 2022) reveals that region feature extraction and hard negative mining are essential for achieving the results reported in their paper but also raise reproducibility concerns. Our approach avoids using region features for simplicity, as they rely on an extra module, and we found that hard negative mining is less efficient in terms of training time. Random in-batch negative contrast works quite well for our proposed dual constraint contrast. Importantly, all the works mentioned above are trained in a supervised manner. In many real-world scenarios involving out-of-domain and downstream tasks, labeled paired data may not be available. To the best of our knowledge, we are the first to propose an unsupervised fine-tuning method that does not require labeled data and achieves new state-of-the-art results compared to supervised baselines.

4.4 Method

In this section, we will first discuss the visual and text embeddings used in our model. Next, to better understand the dual idea, we will explain the prime task, dual task, and cross-modal translation. We will then introduce the architecture of our model and also discuss the skip connection. Lastly, we will discuss the unsupervised dual constraint contrast.

4.4.1 Multimodal embedding and dual task

Visual and text embedding: We opted for a dual encoder (Radford et al., 2021) to achieve fast retrieval performance, which encodes images and text separately. The choice of architecture is flexible, allowing for the use of other fusion encoders (Chen et al., 2020) or multi-way transformer architectures (Bao et al., 2022) if needed. For visual features, we choose grid features extracted from ResNet (K He et al., 2016) and patch projections from vision transformers (Dosovitskiy et al., 2021) as two different visual backbones. Although using region features has been proven to achieve better results in the cross-modal retrieval task, we do not use them as they require additional pre-trained object detection modules like Fast-RCNN (Girshick, 2015) using the Visual Genome dataset (Krishna et al., 2017), which is less efficient. For text features, like most recent works, we utilize BERT embeddings (Devlin et al., 2019). Formally:

$$\begin{aligned} \{v_n\}_{n=1}^{N} &= Encoder_{\text{visual}}(v) \\ \{t_m\}_{m=1}^{M} &= Encoder_{\text{text}}(t) \end{aligned} \tag{4.1}$$

where v and t are the input image and text respectively. Suppose the visual encoder can extract N visual vectors, which can be either grid features or patch projections, in d_1 dimensions. Similarly, the text encoder can extract M token vectors in d_2 dimensions.

After extracting the visual and textual features, we input them into the VLP model to obtain fused representations, which are generalized multimodal representations. Following transformations in $VLP_{vision}(\cdot)$ and $VLP_{text}(\cdot)$, the fused vision features and text features are in the same dimension space, represented as $\mathbb{R}^d$ (i.e. in CLIP $d = 768$).

$$\begin{aligned} \mathbf{v} &= VLP_{\text{vision}}\left(\{\mathbf{v}_n\}_{n=1}^{\text{N}}\right) \\ \mathbf{t} &= VLP_{\text{text}}\left(\{\mathbf{t}_m\}_{m=1}^{\text{M}}\right) \end{aligned} \tag{4.2}$$

Prime and dual task: Text retrieval (image → text) and image retrieval (text → image) are mutually dual tasks. For simplicity and better explanation, we denote the prime modality as the visual modality, and the prime task as text retrieval. In parallel, the dual modality is the text modality, and the dual task is image retrieval. Cross-modal retrieval relies on accurate multimodal alignment and serves as an ideal task to evaluate the performance of multimodal representation learning in terms of **inter-modality** effectiveness.

Prime task - text retrieval: Given a query from the prime modality (vision), we perform text retrieval by measuring the similarity between the query image and candidate texts (dual modality) in the mini-batch as shown in the equation below:

$$\hat{t} = \underset{(t_i, v)\sim\mathcal{B}}{\text{argmax}}\left(\text{sim}\left(t_i, v\right)\right) \tag{4.3}$$

where v is the visual feature of the query image, t_i is the text feature of the candidate texts in the mini-batch $\mathcal{B}$. $\hat{t}$ is the text that is most similar to the query image in the semantic space. The $sim(\cdot)$ function can be cosine similarity or cross-entropy.

Dual task – image retrieval: Likewise, given a query text in the dual modality, we conduct image retrieval by measuring the similarity between query text and candidates images in the prime modality within the mini-batch as the equation below:

$$\hat{v} = \underset{(v_i, t)\sim\mathcal{B}}{\text{argmax}}\left(\text{sim}\left(v_i, t\right)\right) \tag{4.4}$$

where t represents the text feature of the query text, while v_i denotes the visual feature of candidate images within the mini-batch $\mathcal{B}$. $\hat{v}$ corresponds to the image that bears the greatest similarity to the query text within the semantic space.

Cross-modal translation: We introduce a cross-modal translation task, to evaluate **intra-modality** alignment performance. Our preliminary experiments revealed that semantically close instances within unimodal domains, such as visual and language, tend to be separated in the multimodal representation space. This separation can lead to errors like counting mistakes and misclassification of fine-grained features (discussed further in Section 4.6.4). We argue that fused visual and text features should remain close to instances that share similar semantics. To address this, we introduce the cross-modal translation task, which involves using a forward retrieval instance as a candidate to perform a back-retrieval task and then comparing whether the original query in the same modality can still be identified.

More specifically, let's assume the forward-retrieval task as text retrieval. Using Eq. (4.5), we first identify the text instance in the mini-batch that has the maximum

similarity score with the query image. Next, we use this retrieved text candidate to perform a back-retrieval task, image retrieval, finding the image in the same mini-batch with the highest similarity score to the candidate text. Finally, we use a $sim(\cdot)$ function to measure the similarity between the back-retrieved image and the original query image.

$$\text{sim}(v, \hat{v}) = sim(v, \underset{(v_i, \hat{t})\sim\mathcal{B}}{\text{argmax}} \left(\text{sim}\left(v_i, \hat{t}\right)\right)) \tag{4.5}$$

where v represents the visual feature of the query image, while $\hat{v}$ denotes the image obtained from the back-retrieval task. $\hat{t}$ refers to the forward retrieved text candidate, as shown in Eq. (4.3), and v_i represents the image feature in the same mini-batch $\mathcal{B}$.

Similarly, we can initiate the process with the language modality, where the forward-retrieval task is image retrieval and the back-retrieval task is text retrieval, as shown in Eq. (4.6).

$$\text{sim}(t, \hat{t}) = sim(t, \underset{(t_i, \hat{v})\sim\mathcal{B}}{\text{argmax}} (\text{sim}(t_i, \hat{v}))) \tag{4.6}$$

where t denotes the text feature of the query text, while $\hat{t}$ represents the text obtained from the back-retrieval task. $\hat{v}$ refers to the forward retrieved image candidate, as shown in Eq. (4.4), and t_i signifies the text feature in the same mini-batch $\mathcal{B}$.

It is important to note that, since we utilize multimodal representations from VLP models as the backbone, the back-retrieved instances are likely to be the original query. This likelihood is due to the consideration of relevant pairs in the data construction using VLP. Given that VLP multimodal representations are generalized and transferable, and the candidate from the forward retrieval serves as a bridge, we hypothesize that we can leverage this dual process to form a dual constraint contrast loss. This approach would allow the model to be trainable without labeled paired dataset by only adding extra parameters at the output level to learn out-of-domain and task-specific representations. To realize this hypothesis, we introduce a skip connection and unsupervised dual constraint contrast, which will be discussed in the next section.

4.4.2 Framework and skip connection

Given the remarkable generalization and transfer capabilities of VLP models, we use VLP as the backbone and freeze the VLP parameters for parameter-efficient fine-tuning to obtain the fundamental in-domain multimodal representations. For simplicity, we add two linear probe layers at the output level of the VLP backbone to learn out-of-domain and task-specific multimodal representations, as illustrated in Figure 4.2. For visual and text modalities, following Eqs. (4.1) and (4.2), the fused visual and text representations can be expressed as follows:

$$\begin{aligned} v &= FC_v\left(VLP_{\text{vision}}\left(\{v_n\}_{n=1}^{N}\right)\right) \\ t &= FC_t\left(VLP_{\text{text}}\left(\{t_m\}_{m=1}^{M}\right)\right) \end{aligned} \tag{4.7}$$

where $VLP_{\text{vision}}\left(\{v_n\}_{n=1}^{N}\right)$ denotes the fused visual feature from VLP models, and $VLP_{\text{text}}\left(\{t_m\}_{m=1}^{M}\right)$ denotes the fused text feature from VLP models. FC represents linear probe layers. v and t indicate the fused visual and text features after the linear probe layers.

However, since our method is based on unsupervised dual contrast, the model needs to have a basic ability to retrieve candidates as shown in Eqs. (4.5) and (4.6);

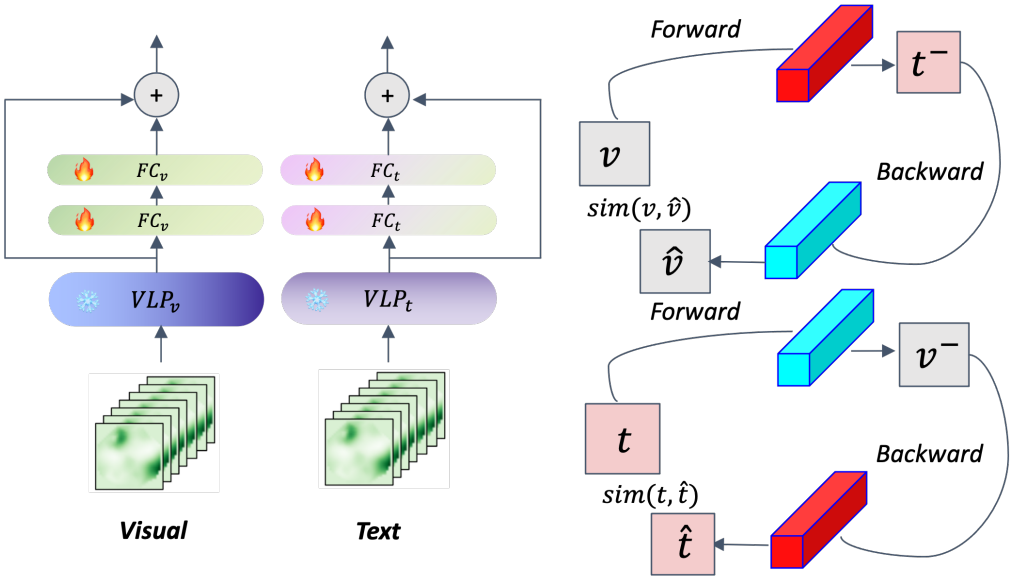


Figure 4.2: Illustration of (left) our framework (Section 4.4.2) and dual constraint contrast (right) (Section 4.4.3).

otherwise, the model will collapse. Therefore, we introduce a skip connection (K He et al., 2016), linking the representation from VLP to the final output prediction of our model as Eq. (4.8). In practice, the skip connection is crucial for robust fine-tuning and fast convergence. This is because, in the beginning, the VLP model will contribute more to retrieving the candidate and allow the method to compare the query and candidate in the same modality.

$$\begin{aligned} v &= \alpha_1 \cdot VLP_{\text{vision}}\left(\{v_n\}_{n=1}^{N}\right) + \alpha_2 \cdot FC_v\left(VLP_{\text{vision}}\left(\{v_n\}_{n=1}^{N}\right)\right) \\ t &= \gamma_1 \cdot VLP_{\text{text}}\left(\{t_m\}_{m=1}^{M}\right) + \gamma_2 \cdot FC_t\left(VLP_{\text{text}}\left(\{t_m\}_{m=1}^{M}\right)\right) \end{aligned} \tag{4.8}$$

where α_1, α_2, γ_1, and γ_2 are hyperparameters to balance the importance of the fused multimodal representations from VLP and the representations after linear probe layers. In experiments, we find that these hyperparameters are not sensitive and are set to 1.0.

Note that our model adds extra parameters at the output level. This approach is more flexible compared to Adapter and Prompt tuning methods. Adapter methods (Houlsby et al., 2019) inject extra parameters at the model level, which require the training code of VLP models, while prompt tuning methods (K Zhou et al., 2022) incorporate and train parameters at the input level, necessitating considerable effort to find the best template. Our model consists of two linear probe layers and a skip connection shortcut. We do not use any complex fusion layers or cross-attention-based methods. As seen in Table 4.2, our methods surpass fine-tuning-based methods and outperform fine-tuned VLP backbones by a large margin. In the following section, we will introduce how to leverage our model to form a dual constraint contrast loss.

4.4.3 Self-Supervised Dual-Constraint Contrastive Learning

As discussed in Section 4.4.1, our dual constraint contrast is formed by the forward retrieval and back retrieval as a loop. More specifically, we first obtain fused image and text features for each mini-batch using Eq. (4.8). Assuming the forward-retrieval task

is text retrieval, we take each image instance v as the query to find the most similar negative text sample t^- in the batch by computing the similarity scores as in Eq. (4.3). In our experiments, we employ the cosine similarity function for measuring the similarity. We refer to the retrieved text as the negative sample since we do not know if it is the anchor in an unsupervised method. We use the forward-retrieved text t^- to conduct the back-retrieval task-image retrieval. Similarly, we compute the similarity scores between the text t^- and all the images in the mini-batch to obtain a similarity vector. We then normalize the similarity vector using the Softmax function. Finally, we form the loss as the cross-entropy loss using the normalized similarity vector with the pseudo-label vector where the original query image is one, and the other images are zero. Likewise, we can begin with the forward retrieval task as the image retrieval task and the back retrieval task as the text retrieval task. We train our model using cross-entropy loss:

$$\mathcal{L}\left(\theta^V, \theta^L\right) = -\frac{1}{|\mathcal{B}|}\left(\sum_{i=1}^{M} y_i^V \log\left(f_{\left(\theta^V,\theta^L\right)}^{V\to L\to V}\left(v, t^-, \hat{v}\right)_i\right) + \sum_{i=1}^{M} y_i^L \log\left(f_{\left(\theta^V,\theta^L\right)}^{L\to V\to L}\left(t, v^-, \hat{t}\right)_i\right)\right) \tag{4.9}$$

where (θ^V, θ^L) are the trainable parameters in the two layers of the linear probe, $|\mathcal{B}|$ is the batch size, and M is the number of instances in the batch. y^V and y^L represent the pseudo-labels where the query image or text is one and other images and text are set to zero. $f^{V\to L\to V}$ represents the loop from text retrieval to image retrieval, and $f^{L\to V\to L}$ is the reverse dual loop, going from image retrieval to text retrieval.

4.5 Experiment Setup

Datasets. We evaluate our method on two widely used benchmark datasets for cross-modal retrieval, MS COCO and Flickr 30K, and one distinct dataset called MOTIF, with more complex text. In more detail, MS COCO contains 123,287 images, each with five sentences describing the image's content. Flickr 30K has 31,783 images; like MS COCO, it is also paired with five corresponding sentences. Following the typical approach to split datasets in most of the literature, we use the Karpathy split (Karpathy and Fei-Fei, 2015) method for MS COCO and Flickr 30K datasets. We use MOTIF, a language-oriented multimodal dataset, to test the domain transfer effect. MOTIF has 1,125 sentences with at least three complex words, and the structure is more complex than the sentences in MS COCO and Flickr 30K. We randomly split the dataset into training and test datasets as 900/225 images. In the implementation within an unsupervised setting, we employ pre-trained VLP models (without exposing any datasets) to conduct cross-modal retrieval. The goal is to find relevant pairs, which may or may not be correct. During training, shuffling is also performed to increase the diversity of negative samples within the mini-batch.

Evaluation metrics. We evaluate the cross-modal retrieval performance and cross-modal translation performance using $recall@K$ as the evaluation metric. In our experiments, we report $R@1$, $R@5$ and $R@10$. To provide a clearer description of the tasks in our experiments, we use the following abbreviations: "IR" for Image Retrieval, "TR" for Text Retrieval, "ITI" for Image-Text-Image Translation, and "TIT" for Text-Image-Text Translation.

Implementation details. We implement our model using the PyTorch framework and utilize the CLIP model as the backbone. The dimension of the fused visual and text features from CLIP is 768. Our model is trained on an Nvidia RTX A6000 GPU, but we only utilize 14GB of its 48GB memory. For optimization, we employ the Adam optimizer with a learning rate of 1e-5 and a weight decay of 1e-5. The balance hyperparameters, α_1, α_2, γ_1, and γ_2, show minimal sensitivity in our experiments and are all set to 1.0.

We use two CLIP architectures as backbones in our experiments. For the text encoder, we employ BERT as the backbone. In the visual encoder, we implement two versions: one using ResNet-50 (K He et al., 2016) and the other using ViT-L/14@336px (Dosovitskiy et al., 2021). To be clear, we refer to the dual contrastive model using ResNet-50 as Successor@RN50 and the one using ViT-L/14@336px as Successor@ViT-L. Both linear probe layers have a dimension of 768 and employ the non-linear activation function, ReLU. We train the Successor@RN50 model for 25 epochs and the Successor@ViT-L model for 20 epochs.

4.6 Results and Analysis

To demonstrate the effectiveness of our proposed method presented in Section 4.4, we compare it with six baselines. First, we compare our method with five pre-trained state-of-the-art methods: ViLBERT (Jiasen Lu et al., 2019), PixelBERT (Z Huang et al., 2020), UNITER (Chen et al., 2020), ViLT (Kim et al., 2021), and CLIP (Radford et al., 2021). Due to reproducibility issues, we cite the results from (Rao et al., 2022) and report and compare them in this chapter. Since CLIP did not use MS COCO and Flickr 30K for pre-training, we adopted the standard linear probing method to fine-tune CLIP. Next, we compare our method with the most recent work on plug-and-play method, BCAR (Diao et al., 2023), in the cross-modal retrieval task. Detailed settings and comparisons can be found in Table 4.1.

Table 4.1: Comparison of our proposed method with five state-of-the-art VLP methods and one plug-and-play method on the image-text retrieval task. For grid features $*$, PixelBERT used ResNet-50 features and CLIP as well as our models used two variants, ResNet-50 and ViT-L patches.

Method	Params	Architecture	Fine-tuning	Visual Tokens	Pre-trained Datasets	BS	Unsupervsied	Loss
ViLBERT	221M	fusion encoder	full fine-tune	Region	CC	64	✗	cross-entropy
PixelBERT	124M	fusion encoder	full fine-tune	Grid*	VG, MSCOCO	512	✗	cross-entropy
UNITER	110M	fusion encoder	full fine-tune	Region	CC, SBU, VG, MSCOCO	64	✗	cross-entropy
ViLT	111M	fusion encoder	full fine-tune	Region	CC, SBU, VG, MSCOCO	256	✗	cross-entropy
CLIP	2.3M	dual encoder	frozen fine-tune	Grid*	WIT	128	✗	contrastive loss
BCAR	2.2M	fusion encoder	frozen fine-tune	Region	VG	128	✗	ranking loss
Successor	2.3M	dual encoder	frozen fine-tune	Grid*	✗	128	✓	cross-entropy

It is worth noting that all the baselines are fine-tuned in a supervised manner, and ViLBERT, UNITER, ViLT, and BCAR use region features, which have been proven to improve results but increasing training time as the involvement extra detector modules. These baselines also employed extra datasets like CC (Sharma et al., 2018), VG (Krishna et al., 2017), and SBU (Ordonez et al., 2011) datasets, as well as other techniques to enhance performance. However, our proposed unsupervised dual contrastive method aims to create a more flexible and adaptable approach for cross-modal retrieval tasks

that do not rely on specific tricks, region features, or a pre-trained object detection module like Fast-RCNN.

By comparing our method with the supervised baselines, we aim to demonstrate the effectiveness of our approach in scenarios where labeled paired datasets are unavailable, and out-of-domain cases. In doing so, we highlight the advantages of our method, which involves lightweight trainable parameters, making it a more practical choice for a wider range of real-world applications.

4.6.1 Comparison to state-of-the-art methods

Table 4.2 presents a comprehensive comparison with state-of-the-art VLP models and one plug-and-play method on Flickr 30K and MS COCO datasets as mentioned above. The top half of the table displays the performance of fine-tuned VLP models trained in a supervised manner. The bottom half showcases the results of our proposed approaches, including the ResNet-50 variant and the ViT variant. The best results are highlighted in blue for the top-performing results among the baseline methods, while red represents the best results achieved by our methods, surpassing the best baseline results.

Table 4.2: Cross-modal retrieval results on MS COCO and Flickr 30K datasets. The top half of the table displays the performance of fine-tuned VLP models using supervised methods, while the bottom half showcases the results of our proposed approaches, including the ResNet-50 variant and the ViT variant. The best results are highlighted in blue and red.

Model	Flickr 30K 1K Test						MS COCO 5K Test					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Supervised VLP Performance												
ViLBERT	58.2	84.9	91.5	76.8	93.7	97.6	38.6	68.2	79.0	53.5	79.7	87.9
PixelBERT	59.8	85.5	91.6	75.7	94.7	97.1	41.1	69.7	80.5	53.4	80.4	88.5
UNITER	62.9	87.2	92.7	78.3	93.3	96.5	37.8	67.3	78.0	52.8	79.7	87.8
ViLT	62.2	87.6	93.2	83.7	97.2	98.1	42.6	72.8	**83.4**	**62.9**	**87.1**	**92.7**
CLIP@RN50	**68.5**	**91.6**	**95.6**	**84.7**	**97.3**	**99.1**	43.1	70.8	80.9	59.7	83.8	90.6
CLIP@ViT-L	**73.7**	**93.2**	**96.3**	**88.3**	**98.7**	**99.5**	**46.5**	**73.4**	82.7	63.6	86.2	92.5
BCAR	62.6	85.8	91.1	82.3	96.0	98.4	44.3	73.2	83.2	61.3	86.1	92.6
Dual Contrast Performance (Ours)												
Successor@RN50	**71.3** ↑	**92.2**↑	**96.0** ↑	**87.6** ↑	**98.5** ↑	**99.3** ↑	43.8	71.4	81.1	60.5	85.1	91.3
Successor@ViT-L	**74.9** ↑	**94.1** ↑	**96.8** ↑	**89.1** ↑	**98.7** ↑	**99.5** ↑	**46.8** ↑	**74.1** ↑	**83.2** -	**64.7** ↑	**86.5** -	**92.7** ↑

Our unsupervised ViT variant outperforms all the baseline results on both Flickr 30K and MS COCO datasets. Additionally, both the ResNet-50 variant and ViT variant surpass all baselines on the Flickr 30K dataset. This demonstrates the effectiveness of our dual constraint contrast methods. In particular, Successor@RN50 achieves a 1.5% (536.8 → 544.9) relative gain, and Successor@ViT-L achieves a 0.62% (549.7 → 553.1) relative gain on the Flickr 30K dataset compared with the best baselines, CLIP@RN50 and CLIP@ViT-L, in supervised VLP performance. Successor@ViT-L also obtains a 1.5% (441.5 → 448) relative gain on MS COCO compared with the best results of the ViLT model. These results highlight the effectiveness of our dual constraint contrast methods and their stability, as our model is simple and omits any tricks for simplicity.

Regarding parameter-efficient fine-tuning performance, we observe that the frozen and fine-tuning paradigm works better than fully fine-tuning the model. From the perspective of trainable parameters, CLIP, BCAR, and our Successor model have 98%

fewer parameters than PixelBERT, UNITER, and ViLT, and 99% fewer parameters than ViLBERT. Nevertheless, CLIP and BCAR achieve the best results among baseline methods on the Flickr 30K dataset and the image retrieval task on the MS COCO dataset, which demonstrates that Parameter-Efficient Fine-Tuning (PEFT) methods are more efficient for fine-tuning while maintaining high accuracy. ViLT attains the best results in MS COCO as the dataset is much larger, and learning from supervised labels helps improve accuracy. Importantly, our methods outperform all the PEFT baselines and achieve better or comparable results to all the baselines and even the ViLT model on MS COCO.

Lastly, compared with fine-tuned CLIP, our Successor shares the same architecture but achieves similar or even better results on both datasets in an unsupervised manner. This supports our hypothesis that the VLP model possesses exceptional generalization and capabilities. We can fine-tune the model by adding extra layers at the output level to inherit the abilities of VLP and learn out-of-domain and task-specific representations without labeled data.

The improved performance of our method demonstrates that the learned out-of-domain and task-specific multimodal representations possess strong inter-modality effectiveness. As discussed in Section 4.4.1, cross-modal translation tasks can evaluate the intra-modality alignment of multimodal representations. In Table 4.3, we conduct cross-modal translation and compare Successor@RN50 and Successor@ViT-L with the baseline CLIP@RN50 and CLIP@ViT-L on Flickr 30K and MS COCO datasets. Both variants consistently achieve better results than the baseline, illustrating the effectiveness of our method. As we opted for CLIP as the VLP backbone, the improved results indicate that closely related semantic instances within a unimodal representation maintain their proximity in the multimodal representation space compared with VLP models. This highlights the successful intra-modality alignment achieved by our dual constraint contrast methods.

Table 4.3: Cross-modal translation results on MS COCO and Flickr 30K datasets. The **bold** number represents the best results achieved by models.

Model	Flickr 30K 1K Test						MS COCO 5K Test					
	ITI@1	ITI@5	ITI@10	TIT@1	TIT@5	TIT@10	ITI@1	ITI@5	ITI@10	TIT@1	TIT@5	TIT@10
CLIP@RN50	87.9	99.7	100.0	65.0	82.9	93.6	71.2	98.0	99.6	40.0	67.3	82.2
CLIP@ViT-L	91.0	99.9	100.0	70.9	**86.4**	94.8	73.1	97.8	99.7	43.2	68.1	81.9
Successor@RN50	91.2	**100.0**	100.0	68.5	84.4	92.7	72.6	97.8	99.7	40.3	67.6	82.2
Successor@ViT-L	**92.0**	99.8	**100.0**	**71.8**	86.2	**95.1**	**74.3**	**98.3**	**99.8**	**43.8**	**68.8**	**83.1**

4.6.2 Zero-shot performance

The multimodal representations obtained from VLP models have demonstrated remarkable generalization capabilities. Our proposed method involves freezing the VLP as the foundation and adding linear probe layers at the output level. Thus, training the model with extra data should act as the incremental of the foundation knowledge of VLP. We hypothesize that this approach should yield better zero-shot performance compared to the original VLP. To test this hypothesis, we first train our proposed model described in Section 4.4 using the dual constraint contrast method on the Flickr 30K dataset and evaluate the model on the MS COCO 5K test set. Similarly, we train our proposed model on the MS COCO dataset and test it on the Flickr 30K dataset. As

demonstrated in Tables 4.4 and 4.5, our fine-tuned model with the dual contrast method achieves better or comparable zero-shot performance compared to the fine-tuned CLIP model with the same backbone on both datasets. We attribute these improvements to the frozen and fine-tuning paradigm. In comparison to the full fine-tuning approach, full fine-tuned models run the risk of causing catastrophic forgetting (J Li et al., 2022), where the previously learned generalized and transferable multimodal representations from VLP models degrade. By using the frozen and fine-tuned paradigm, we can avoid this issue and maintain the quality of multimodal representations, leading to improved zero-shot performance.

Table 4.4: Zero-shot performance results on the Flickr 30K dataset. The **bold** number represents the best results achieved by models.

Model	Flickr30K-IR			Flickr30K-TR		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP@RN50	61.5	84.7	90.0	81.8	95.9	98.1
CLIP@ViT-L	64.0	86.6	91.6	84.8	**97.9**	**99.1**
Successor@RN50	66.9	89.2	93.2	82.7	97.0	98.6
Successor@ViT-L	**70.6**	**91.7**	**95.1**	**86.5**	97.4	98.9

Table 4.5: Zero-shot performance results on the MS COCO dataset. The **bold** number represents the best results achieved by models.

Model	MSCOCO5K-IR			MSCOCO5K-TR		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP@RN50	35.1	59.7	69.9	54.8	78.8	86.4
CLIP@ViT-L	36.8	61.4	71.3	57.5	81.1	87.7
Successor@RN50	38.2	64.0	74.0	55.6	78.9	86.5
Successor@ViT-L	**42.7**	**68.1**	**77.6**	**60.2**	**82.0**	**89.4**

4.6.3 Domain adaptation performance

To better investigate the domain adaptation performance of our proposed model, we train and compare our method with the baseline CLIP model on the MOTIF dataset. The MOTIF dataset is an education-oriented multimodal dataset, where the sentence structure and vocabulary are more complex than those in the Flickr 30K or MS COCO datasets. Table 4.6 demonstrates that our proposed method performs well in acquiring out-of-domain multimodal representations compared to the supervised method on CLIP. In addition to its unsupervised attributes, our proposed model can effectively train and transfer knowledge on a small dataset without overfitting. This characteristic makes it particularly useful for domain adaptation tasks, where it is essential to leverage and adapt existing knowledge to new, complex domains with limited labeled data available.

Table 4.6: Domain adaptation performance results on the MOTIF dataset. The **bold** number represents the best results achieved by models.

Model	MOTIF-IR			MOTIF-TR		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP@RN50	48.0	96.8	**100.0**	48.0	81.6	90.4
Successor@RN50	**50.4**	**97.6**	99.2	**48.0**	**82.4**	**92.8**

Table 4.7: Ablation study results on Flickr 30K dataset. The underlined number represents a degradation in performance.

Model	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
Successor@RN50	**71.3**	**92.2**	**96.0**	**87.6**	**98.5**	**99.3**
- $V \rightarrow L \rightarrow V$	63.5	87.3	92.6	85.5	97.6	99.1
- $L \rightarrow V \rightarrow L$	70.1	91.3	95.5	79.5	94.7	97.9
Successor@ViT-L	**74.9**	**94.1**	**96.8**	**89.1**	**98.7**	**99.5**
- $V \rightarrow L \rightarrow V$	65.0	87.4	93.1	88.0	98.2	99.5
- $L \rightarrow V \rightarrow L$	74.1	93.4	96.4	84.6	97.0	99.1

4.6.4 Error analysis and ablation study

In addition to quantitative analysis, we also examine the quality of retrieved results through a qualitative investigation. We observe that VLP models sometimes exhibit fine-grained errors, primarily related to counting mistakes, nuanced color and pattern understanding, and complex noun and verb comprehension. For instance, the term "runners" is the plural form of "runner"; however, the CLIP model overlooks this vital information. Our proposed model is capable of capturing fine-grained multimodal representations. Meanwhile, we conduct ablation studies using two variants on the Flickr 30K dataset. For each model, we remove either the dual constraint loss from $V \rightarrow L \rightarrow V$ or $L \rightarrow V \rightarrow L$. The results in Table 4.7 show that when the loss from $V \rightarrow L \rightarrow V$ is removed, the image retrieval performance degrades, while removing the loss from $L \rightarrow V \rightarrow L$ causes the text retrieval performance to degrade.

4.7 Conclusion

In this work, we present an unsupervised dual constraint contrast method designed to efficiently fine-tune VLP models using a "frozen and fine-tuning" paradigm. By incorporating additional linear probe layers at the output level and incorporating a skip shortcut, we achieve fast convergence. As our approach only updates lightweight parameters (2.3M), the training cost is significantly lower compared to other full fine-tuning methods. Consequently, we can train our model using a single GPU, achieving convergence within hours while maintaining comparable or superior performance to existing fine-tuned VLP and PEFT methods on two benchmark datasets. Furthermore, our method demonstrates strong domain transfer capabilities. With its simplicity and feasibility, our approach is agnostic to the underling models and has the potential to harness the power of more advanced VLP models in the future.

The first principle is that you must not fool yourself.

— Richard P. Feynman (1974)

5

Mitigating Hallucinations with Instruction Contrastive Decoding

Contents

Publication Note. This chapter is based on a first-authored publication by the dissertation author: Wang et al., "Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding," published in Findings of the Association for Computational Linguistics (ACL 2024). For integration into this dissertation, the material has been lightly revised for consistency in terminology, formatting, and cross-references. The core contributions and findings remain unchanged.

5.1 Abstract

Large Vision-Language Models (LVLMs) are increasingly adept at generating contextually detailed and coherent responses from visual inputs. However, their application in multimodal decision-making and open-ended generation is hindered by a notable rate of hallucinations, where generated text inaccurately represents the visual contents. To address this issue, this study introduces the Instruction Contrastive Decoding (ICD) method, a novel approach designed to reduce hallucinations during LVLM inference. Our method is inspired by our observation that what we call disturbance instructions significantly exacerbate hallucinations in multimodal fusion modules. ICD contrasts distributions from standard and instruction disturbance, thereby increasing alignment uncertainty and effectively subtracting hallucinated concepts from the original distribution. Through comprehensive experiments on discriminative benchmarks (POPE and MME) and a generative benchmark (LLaVa-Bench), we demonstrate that ICD significantly mitigates both object-level and attribute-level hallucinations. Moreover, our method not only addresses hallucinations but also significantly enhances the general perception and recognition capabilities of LVLMs.

5.2 Introduction

Recent research in large vision-language models (LVLMs) (Haotian Liu et al., 2023; Haotian Liu et al., 2024; J Li et al., 2023) has seen remarkable progress, benefiting from the integration of advanced large language models (LLMs) (Achiam et al., 2023; Touvron, Lavril, et al., 2023; Touvron, Martin, et al., 2023) known for their robust language generation and zero-shot transfer capabilities (Zhong et al., 2023; Peng et al., 2023). In order to leverage off-the-shell LLMs, it is crucial to facilitate cross-modal alignment. LLaVa (Haotian Liu et al., 2023) employs a linear projection approach, while BLIP-2 (J Li et al., 2023) and InstructBLIP (Haotian Liu et al., 2024) narrow the modality gap using a Q-Former. Although LVLMs have shown promising outcomes, the issue of hallucination remains. This phenomenon occurs when the generated textual content, despite being fluent and coherent, does not accurately reflect the factual visual content.

The object hallucination was initially explored within the realm of image captioning (Rohrbach et al., 2018). As LVLMs harness the sophisticated understanding and generative prowess of LLMs, the scope of hallucination extends beyond mere object existence. It now encompasses more complex elements such as attributes and relationships within the generated content. Consequently, distinguishing discriminative hallucination and

the non-hallucinatory portion in the generation has become pivotal in assessing the performance of LVLMs in terms of their fidelity to factual visual information.

The intertwined nature of modalities presents significant challenges in identifying the root causes of hallucinations in LVLMs. Research efforts have begun to uncover the primary contributors to LVLM hallucinations, including statistical biases (You et al., 2024) encountered during the training process and excessive dependence on language priors (X Zhu et al., 2020; Zhibo et al., 2023). Additionally, multimodal misalignment has been identified as a key factor in the occurrence of hallucinations (C Jiang et al., 2024; F Liu et al., 2024; F Wang et al., 2024). To address dataset bias, annotation enrichment techniques (Gunjal et al., 2024; You et al., 2024; Zhai et al., 2024) have been introduced. Furthermore, to counteract the influence of language priors, post-processing strategies (S Yin et al., 2024; Y Zhou et al., 2024) have been developed, along with comprehensive initiatives aimed at improving multimodal alignment through optimizing alignment with humans (Sun et al., 2024; C Jiang et al., 2024). While these interventions have proven to be effective in reducing hallucinations, they demand substantial human involvement and incur significant computational costs for additional training or the integration of supplementary modules.

In this work, we reveal that appending instructions with role prefixes to form disturbance instructions can significantly exacerbate hallucinations. We hypothesize that identifying and subsequently detaching hallucination concepts from the original distribution could effectively reduce such hallucinations. Motivated by this insight, we introduce the **Instruction Contrastive Decoding** (**ICD**) method. This approach is novel in that it is training-free and agnostic to the underlying LVLMs. ICD differentiates between two distributions: one from the original instruction and another from the disturbance instruction within the multimodal alignment module. Utilizing their difference, we aim at suppressing hallucinations. Through comprehensive experiments on discrimination hallucination benchmarks such as POPE (Y Li et al., 2023) and MME hallucination sets (Fu et al., 2025), as well as the generation hallucination benchmark LLaVa-Bench (Haotian Liu et al., 2023), our method incorporating state-of-the-art LVLMs like miniGPT4 and InstructBLIP, demonstrates significant efficacy in mitigating hallucinations at both object and attribute levels. Furthermore, our approach consistently enhances performance across 14 general perception and recognition tasks within the full MME benchmark.

Our main **contributions** are as follows:

- We perform an in-depth analysis of how disturbance in instructions exacerbates hallucinations. This phenomenon is elucidated through statistical bias and language priors, offering a nuanced understanding of underlying causes.

- Drawing on these insights above, we introduce the ICD method. This novel strategy, which emphasizes initial highlight followed by de-emphasize of hallucination, effectively mitigates hallucinations during inference, by adjusting the distributions away from hallucinations that we elicit.

- Through extensive experimentation and analysis, we validate the effectiveness of our proposed ICD method across both discrimination and generation hallucination benchmarks, showcasing its robustness and versatility in enhancing LVLMs performance.

5.3 Related Work

Large Vision-Language Models. The field of vision-language pre-training (VLP) (Radford et al., 2021; J Li et al., 2022; Bao et al., 2022; W Wang et al., 2023) and fine-tuning (**Wang et al.**, 2023; Wiehe et al., 2022; Alayrac et al., 2022) have seen rapid advancements, propelled by the evolution of large language models (LLMs). As a result, large vision-language models (LVLMs) have emerged, leveraging the strengths of frozen LLMs while emphasizing the facilitating of multimodal alignment modules. Notably, models such as LLaVa and Qwen-VL (J Bai et al., 2023) adopt simple linear projections to achieve alignment, contrasting with BLIP-2 and miniGPT4 (D Zhu et al., 2024), which introduce a Q-Former. In further work, InstructBLIP integrates task-aware instructions, enriching the understanding of task-aware visual semantics. Our research builds upon these advancements in LVLMs, focusing on the impact of instruction disturbances. We explore how such disturbances increase the uncertainty in multimodal alignment, significantly contributing to the exacerbation of hallucinations.

Hallucination in VLMs. Generation hallucinations have been extensively studied in the field of language modeling (Tonmoy et al., 2024; Wen et al., 2024). Hallucination in VLMs manifests as detailed, fluent, and coherent responses that inaccurately reflect the visual context, including erroneous objects, attributes, and relations (Hanchao Liu et al., 2024; Jing et al., 2024). Various strategies have been proposed to curb hallucinations. Annotation enrichment techniques like M-HalDetect (Gunjal et al., 2024) and GRIT (You et al., 2024), as well as approaches such as HACL (C Jiang et al., 2024) and LLaVA-RLHL (F Liu et al., 2024), seek to improve alignment with human instructions through additional annotations. Similarly, Woodpecker (S Yin et al., 2024) introduces post-processing aimed at mitigating biases from language priors. While these methods have shown promise in reducing hallucinations, they often require extensive data annotation, fine-tuning, and supplementary modules, complicating their implementation. In contrast, our method directly addresses hallucinations during inference. Additionally, Leng et al. (2024) introduced a visual contrastive decoding (VCD) approach that contrasts with the distributions of distorted visual inputs, a concept that bears resemblance to our method. However, our ICD method suppresses hallucinations through disturbance instructions affecting multimodal alignment.

5.4 Method

5.4.1 Inference in LVLMs

Large Vision-Language Models (LVLMs) are comprised of three pivotal components: a visual encoder, a fusion module, and a language model. For processing an input image, a pre-trained visual encoder, such as ViT-L/14 from CLIP (Radford et al., 2021), is employed to extract visual features, denoted as $\mathbf{X_V}$. The fusion module facilitates multimodal alignment. For instance, InstructBLIP introduces an instruction-aware querying transformer. Q-Former, a lightweight transformer architecture, utilizes K learnable query vectors $\mathbf{Q_K}$ to refine the extraction of visual features, thereby enhancing multimodal alignment. It allows the instruction $\mathbf{X_{ins}}$ to interact with the query vectors, fostering the extraction of task-relevant image features:

$$Z_V = Q_\theta(X_V, Q_K, X_{ins}), \tag{5.1}$$

where, $Z_V = Q_\theta(\cdot)$ represents the fused visual features, conditioned on the instructions. Given its sophistication and effectiveness in multimodal alignment, we advocate for the adoption of the instruction-aware Q-Former architecture.

For text queries $\mathbf{X_q}$, a large language model, parameterized by ϕ, such as Vicuna (Chiang et al., 2023), processes the query, leveraging the derived visual features to formulate responses:

$$Y_R = LLM_\phi(H_V, X_{ins}), \tag{5.2}$$

where $H_V = g(Z_V)$ is the transformation ensuring the same dimensionality as the word embedding of the language model. By default, the instruction is the same as text query for both Q-Former and LLM as $\mathbf{X_{ins}} = \mathbf{X_q}$.

Mathematically, in the decoding phase, the response $\mathbf{R}$ can be defined as a sequence of length $\mathbf{L}$, sampled from a probability distribution:

$$p(Y_R|X_V, X_q) = \prod_{t=1}^{L} p_\phi(y_t|H_V, X_q, y_{<t}), \tag{5.3}$$

where $y_{<t}$ represents the sequence of generated tokens up to the time steps $(t-1)$. In the decoding phase of LVLMs, hallucinations often emerge when probable tokens lack grounding in the visual context. C Jiang et al. (2024) and F Liu et al. (2024) indicate that multimodal misalignment is a critical factor contributing to the generation of hallucinations. Thus, we conduct an in-depth analysis of the fusion module, specifically focusing on multimodal alignment. Our work first demonstrates that instructions within the multimodal alignment module can exacerbate hallucinations. To address this, we introduce instruction disturbance and propose an instruction contrastive decoding method, employing a **highlight and then detach** strategy.

5.4.2 Instruction Can Amplify Hallucination

Prior studies have attributed the occurrence of hallucinations in LVLMs to statistical biases within multimodal training datasets (You et al., 2024) and an over-reliance on language priors (X Zhu et al., 2020; Zhibo et al., 2023). Extending this line of observation, we introduce the concept of instruction disturbance in this section. A prefix appended to instructions affects multimodal alignment, thereby exacerbating statistical biases and the over-reliance on language priors.

Introduction of instruction disturbance: We introduce the concept of instruction disturbance, which entails appending a *role prefix* to the original instructions delineated in Section 5.4.1. This disturbance aims to modulate the multimodal alignment uncertainty within LVLMs. As illustrated in Figure 5.1, the base instruction *"Describe this photo in detail"* is combined with learned query vectors in the Q-Former. To implement instruction disturbance, we append either *positive* or *negative* prefixes to the base instruction. Positive prefixes aim to increase the LVLM's confidence in multimodal alignment. Conversely, negative prefixes are designed to reduce the model's alignment confidence.

$$X_{ins} = \begin{cases} [X_d, X_q] & \textit{if disturbance} \\ X_q & \textit{otherwise} \end{cases}, \tag{5.4}$$

where X_d denotes the role prefix, and X_q represents the original instruction. Through this method, we strategically influence the LVLM's confidence level in multimodal alignment by either encouraging a more definitive understanding or introducing ambiguity.

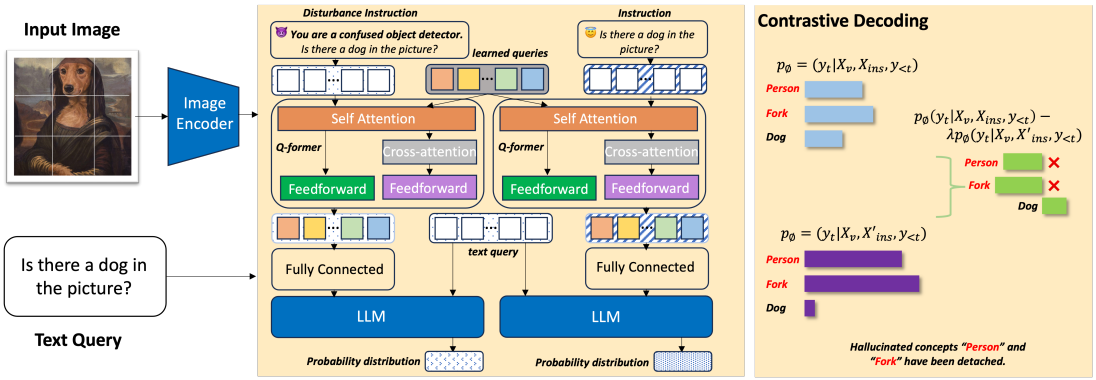


Figure 5.1: An illustration on inference framework and contrastive decoding process of ICD method. At the core (middle orange box), the framework integrates a frozen image encoder, LLM, and query vectors (gray box) within the Q-Former, focusing solely on adjusting the standard and disturbance instructions. The latter, exemplified by adding role prefixes like '**You are a confused object detector**,' aims to increase multimodal alignment uncertainty. This results in two distinct distributions: one from the standard instruction and another influenced by the disturbance. The contrastive decoding method (right orange box) highlights how disturbance instructions amplify hallucinated concepts ('person and fork'), which are then corrected by subtracting probabilities derived from the standard instruction, ensuring accurate recognition of the correct concept 'dog'.

Instruction disturbance amplifies statistical biases and language priors: Figure 5.1 presents the response from InstructBLIP, revealing that the LVLMs generate hallucinated tokens such as "*fork and person*." To further explore this phenomenon, we undertake two specific analyses: the frequent hallucinated object occurrence and the co-occurrence of object hallucinations. Our study utilizes the MS COCO validation set (Lin et al., 2014), a common dataset for LVLM pre-training, to perform hallucination detection across three distinct scenarios: the baseline LVLM, LVLM with a positive disturbance, and LVLM with a negative disturbance. Our analysis focuses on calculating the hallucination ratio, specifically identifying instances where the hallucinated objects are absent from the provided images.

Figure 5.2 demonstrates that introducing instruction disturbance significantly amplifies the occurrence of hallucinations. Under the influence of negative disturbance, LVLMs are more likely to hallucinate objects that frequently co-occur, such as "*person and dining table*," and show an increased tendency to hallucinate objects that typically co-occur with those actually present in the image, for example, "*fork and person*." This suggests that instruction disturbances, whether positive or negative, intensify the hallucination effect, exacerbating the issues of imbalanced object distribution and correlation patterns inherent in the training dataset.

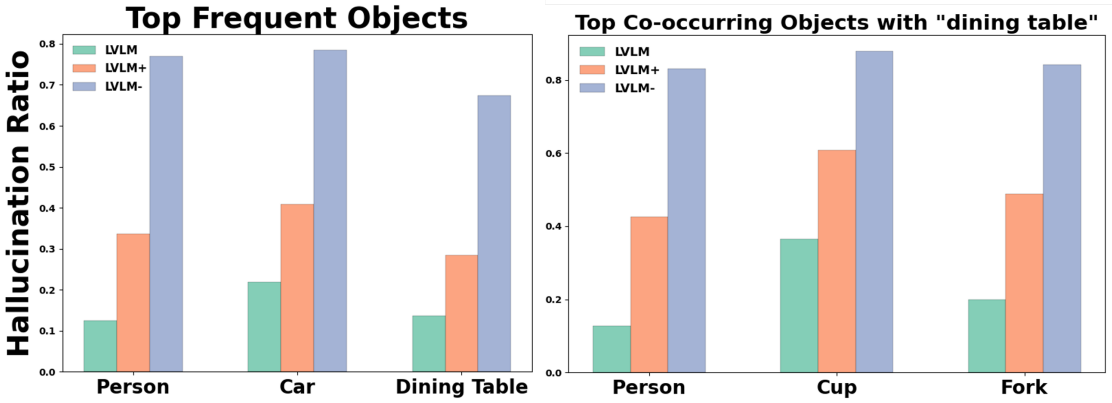


Figure 5.2: The left figure shows the **top frequent objects hallucination ratio** and the right depicts the **ratio of co-occurring object hallucinations** with *dining table*.

5.4.3 Instruction Contrastive Decoding

5.4.3.1 Contrastive Decoding with Disturbance

Our analysis reveals that instruction disturbances exacerbate hallucinations by increasing multimodal alignment uncertainty. This uncertainty predisposes LVLMs to more readily adopt biased co-occurrence concepts from pretraining datasets, as reflected in the learned query vectors. As these hallucinations accumulate, LVLMs increasingly over-rely on language priors. Notably, disturbances involving negative prefixes significantly intensify these hallucinations. We hypothesize that by initially emphasizing the probabilities of hallucinated concepts and subsequently detaching these from the original probability distribution, hallucinations may be reduced. Inspired by this insight, we introduce an Instruction Contrastive Decoding method (ICD) aimed at mitigating hallucinations during LVLM inference.

Motivated by the language contrastive decoding (Sennrich et al., 2024) in reducing hallucinations within machine translation frameworks—where it prevents potentially accurate translations that, however, deviate from the desired target language—we adopt a similar approach to our model. Given the extraction of visual features $\mathbf{X_V}$ from the visual encoder and a textual query $\mathbf{X_q}$, our model calculates two distinct token distributions: one conditioned on the original instructions, and the other on instructions with disturbance $\mathbf{X_d}$ as Equation 5.4. Contrary to the conventional approach of selecting the token that maximizes the probability, our strategy involves choosing the token that concurrently maximizes $p_\phi(y_t|X_V, X_{ins})$ and minimizes $p_\phi(y_t|X_V, X'_{ins})$, the latter representing the probability of tokens that are more likely to be hallucinations. To adjust the balance between these probabilities, we introduce a hyperparameter λ, which regulates the intensity of the contrastive penalty. Formally, this process is described as follows:

$$p_{icd}(Y_R|X_V, X_q) = \prod_{t=1}^{L} \Big(p_\phi(y_t|X_V, X_{ins}, y_{<t}) - \lambda p_\phi(y_t|X_V, X'_{ins}, y_{<t}) \Big), \tag{5.5}$$

where larger λ indicates a more decisive penalty on the decision made by LVLMs with disturbances.

5.4.3.2 Adaptive Plausibility Constraints

The ICD objective is designed to favor tokens preferred by the LVLM output while imposing penalties on tokens influenced by instruction disturbances. However, this approach might inadvertently penalize accurate predictions—those tokens that, under both standard and disturbance instruction conditions, are confidently identified and are well-grounded in the visual context (such as objects, verbs, attributes, and relations) due to their simplicity and high likelihood. Conversely, it might erroneously reward tokens representing implausible concepts. To address this issue, we draw inspiration from adaptive plausibility constraints utilized in open-ended text generation (XL Li et al., 2023). Consequently, we refine the ICD objective to incorporate an adaptive plausibility constraint:

$$\begin{aligned} y_t \sim softmax\Big(& logit_\phi(y_t|X_V, X_{ins}, y_{<t}) \\ & - \lambda logit_\phi(y_t|X_V, X'_{ins}, y_{<t}) \Big) \\ & subject\ to\ y_t \in \mathcal{V}_{head}(y_{<t}) \end{aligned} \tag{5.6}$$

$$\begin{aligned} \mathcal{V}_{head}(y_{<t}) = \Big\{ & y_t \in \mathcal{V} \ : \ p_\phi(y_t|X_V, X_{ins}, y_{<t}) \\ & \geq \alpha \max_{token} p_\phi(token|X_V, X_{ins}, y_{<t}) \Big\}, \end{aligned} \tag{5.7}$$

here, α acts as a pivotal hyperparameter that modulates the truncation of the probability distribution, effectively tailoring the LVLM's response to its confidence level. This is particularly crucial for mitigating the influence of implausible tokens, especially when LVLMs exhibit high confidence and are accurately anchored in visual semantics.

ICD serves as a self-corrective mechanism, which successfully identifies hallucinations in LVLMs and then de-emphasizes them through contrastive decoding. Moreover, the integration of adaptive plausibility constraints further hones the contrastive distribution by considering the confidence levels of LVLMs, thereby narrowing the decision-making process to a more reliable candidate pool. This method not only significantly reduces hallucinations within LVLMs but also curtails the generation of implausible tokens, showcasing the efficacy of our proposed method in enhancing model reliability and output validity.

5.5 Experiment

In this section, we explore the evaluation of our ICD method for mitigating hallucinations. Our examination is twofold: firstly, through the lens of hallucination discrimination, and secondly, via the generation of non-hallucinatory content. More precisely, we assess the efficacy of ICD in alleviating object-level hallucination symptoms utilizing the POPE benchmark. Furthermore, we extend our analysis to include both object and attribute-level symptoms through the MME benchmark. Finally, the performance of our method in generating non-hallucinatory content is evaluated using the LLaVa-Bench dataset.

5.5.1 Experimental Settings

5.5.1.1 Datasets and Evaluation Metrics

POPE: The Polling-based Object Probing Evaluation (POPE) stands as a popular benchmark in discerning hallucination at the object level. POPE employs a binary question-answering format, inquiring LVLMs to determine the presence or absence of a specified object within a given image. This benchmark is structured around three distinct subsets—MS COCO, A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019)—each comprising 500 images alongside six questions per image. POPE introduces three settings within each subset: *random* (selecting absent objects at random), *popular* (choosing the most frequently occurring objects in the dataset as absent), and *adversarial* (selecting absent objects that often co-occur with ground-truth objects). We adopt Accuracy, Precision, Recall, and F1 score as the evaluation metrics.

MME: MME benchmark serves as a comprehensive tool for assessing the capabilities of LVLMs across both perception and cognition, spanning a total of 14 tasks. Among these, tasks focusing on *existence, count, position, and color* are specifically designed as hallucination discrimination benchmarks. These tasks aim to scrutinize both *object-level* and *attribute-level* hallucination symptoms. MME similarly utilizes a question-answering format to facilitate this evaluation. Consequently, task scores are reported as the evaluation metric for measuring performance.

LLaVa-Bench: The LLaVa-Bench is designed to quantify the extent of hallucinated content produced during the open-ended generation tasks performed by LVLMs. This benchmark encompasses a varied collection of 24 images, accompanied by 60 questions that cover a wide range of scenarios, including indoor and outdoor scenes, memes, paintings, and sketches. Unlike discriminative benchmarks, where accuracy serves as the evaluation metric, generative benchmarks, such as this, currently do not have well-established metrics specifically devised for the detailed analysis of hallucinations (Hanchao Liu et al., 2024). Therefore, we utilize case studies on this dataset as a means to qualitatively evaluate the effectiveness of our ICD method (see Section 5.5.5).

5.5.1.2 LVLM Baselines

We employ two state-of-the-art LVLMs as backbone frameworks. Specifically, we implement our ICD on InstructBLIP and miniGPT4, which utilize the Vicuna 7B as their underlying LLM and the sophisticated Q-Former architecture for fusion modules, respectively. Additionally, we explore the use of LLaVa-1.5 (Haotian Liu et al., 2024), which incorporates linear projection for its fusion module alongside InstructBLIP, to identify optimal practices in applying the ICD method (see Section 5.5.4). Finally, we compare our method against the visual contrastive decoding approach (Leng et al., 2024), designed to mitigate hallucinations arising from visual uncertainties. We posit that our method, being LVLM-agnostic, can be conveniently integrated into various off-the-shelf LVLMs.

5.5.1.3 Implementation Details

In our experiments, we adopted the contrastive decoding configurations by setting the decisive penalty on the decision made by LVLMs with disturbance $\lambda = 1$ and the hyperparameter $\alpha = 0.1$ that modulates the truncation of the probability distribution,

in line with the configurations reported in previous studies (XL Li et al., 2023; Leng et al., 2024). For the decoding strategy, we uniformly applied the sampling method across all experiments, incorporating a $top\ p = 1$, a $repetition\ penalty = 1$, and a $number\ of\ beams = 1$ for LLMs. For both VCD and ICD methods, we sample from the modified $softmax$ distribution, as delineated in Equation 5.7.

We conducted experiments with various instructional disturbances, incorporating both positive and negative role prefixes. For illustrative purposes, we present two positive role prefix instructions and two negative role prefix instructions, providing a detailed guide for others to effectively implement our method in practical applications.

- □ P1: *You are an object detector to recognize every different object.*
- □ P2: *You are an object detector to recognize every different object by focusing on the shapes, colors, and relationships of objects.*
- △ N1: *I want you to avoid any specific identification or categorization of the objects depicted.*
- △ N2: *You are a confused object detector to provide a fuzzy overview or impression of the image.*

5.5.2 Experimental Results

5.5.2.1 Results on POPE

The experimental results on POPE, summarized in Table 5.1, demonstrate the efficacy of our instruction contrastive decoding method across three distinct subsets within the POPE benchmark—MS COCO, A-OKVQA, and GQA settings. Notably, our ICD method consistently outperforms the foundational LVLMs, miniGPT4, and InstructBLIP. Specifically, the ICD method exceeds the performance of miniGPT4 and InstructBLIP, showing a substantial improvement of **10.5%** and **6.0%**, respectively, across all metrics (**7.0%** in accuracy, **8.5%** in precision, **8.7%** in recall, and **7.9%** in F1 score for both models). This significant enhancement as per four metrics on POPE underscores the effectiveness of our *highlight and then detach* strategy.

Furthermore, the progressive movement from *random* to *popular* and then to *adversarial* settings reveals a marked decline in performance, highlighting the growing impact of statistical biases and language prior to contributing to hallucinations in LVLMs. Despite these challenges, our ICD method consistently demonstrates improvements across all settings, affirming our hypothesis that disturbance instruction exacerbates hallucinations by influencing multimodal alignment, thereby deepening errors rooted in statistical bias and over-reliance on language priors, which can be subtracted by contrastive decoding. Our method effectively mitigates these issues and object-level hallucinations.

In comparison to the VCD approach, our ICD method achieves an overall improvement of **3.9%**. While the VCD method aims to ensure that the output distributions are closely aligned with visual inputs and compares distributions derived from distorted images, it requires additional processing to distort images via diffusion models (Ho and Salimans, 2021) and is sensitive to the choice of hyperparameters in its experimental

Dataset	Setting	Method	miniGPT4 Backbone				InstructBLIP Backbone			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MSCOCO	Random	*default*	67.04	69.06	66.54	67.77	80.71	81.67	79.19	80.41
		+vcd	69.60	72.76	66.73	69.62	84.53	88.55	79.32	83.68
		+icd	**73.51**	**74.36**	**76.87**	**75.60**	**86.43**	**92.01**	**80.73**	**85.61**
	Popular	*default*	60.89	61.34	65.74	63.46	78.22	77.87	78.85	78.36
		+vcd	62.91	63.69	64.81	64.24	81.47	82.89	79.32	81.07
		+icd	**67.61**	**66.69**	**76.87**	**71.42**	**82.93**	**84.45**	**80.73**	**82.55**
	Adversarial	*default*	59.42	59.64	64.45	61.95	75.84	74.30	79.03	76.59
		+vcd	62.07	62.15	66.76	64.37	79.56	79.67	79.39	79.52
		+icd	**64.36**	**63.68**	**75.11**	**68.93**	**80.87**	**80.95**	**80.73**	**80.84**
A-OKVQA	Random	*default*	64.79	65.26	65.73	65.50	80.91	77.97	86.16	81.86
		+vcd	66.68	66.47	68.21	67.33	84.11	82.21	87.05	84.56
		+icd	**69.04**	**68.50**	**77.04**	**72.52**	**85.82**	**83.80**	**88.94**	**86.29**
	Popular	*default*	60.75	60.67	68.84	64.50	76.19	72.16	85.28	78.17
		+vcd	62.22	**62.23**	68.55	65.24	79.78	76.00	87.05	81.15
		+icd	**62.81**	61.62	**75.78**	**67.97**	**81.64**	**78.50**	**88.77**	**83.32**
	Adversarial	*default*	58.88	58.56	68.50	63.14	70.71	65.91	85.83	75.56
		+vcd	60.67	**60.56**	68.47	64.28	74.33	69.46	86.87	77.19
		+icd	**60.71**	59.27	**77.68**	**67.24**	**74.42**	**70.24**	**88.93**	**78.48**
GQA	Random	*default*	65.13	65.38	66.77	66.07	79.75	77.14	84.29	80.56
		+vcd	67.08	68.30	69.04	68.67	83.69	81.84	**86.61**	84.16
		+icd	**72.24**	**75.08**	**79.54**	**77.24**	**85.10**	**84.21**	86.40	**85.29**
	Popular	*default*	57.19	58.55	60.81	59.66	73.87	60.63	84.69	76.42
		+vcd	62.14	**61.14**	72.26	66.24	78.57	74.62	86.61	80.17
		+icd	**62.84**	61.09	**80.54**	**69.48**	**78.80**	**75.15**	**87.53**	**80.87**
	Adversarial	*default*	56.75	56.26	67.99	61.57	70.56	66.12	84.33	74.12
		+vcd	57.78	57.70	69.82	63.18	75.08	70.59	85.99	77.53
		+icd	**59.64**	**58.21**	**76.81**	**66.23**	**75.17**	**70.59**	**86.27**	**77.65**

Table 5.1: Results on discrimination hallucination benchmark POPE. The default under methods denotes the standard decoding, whereas VCD represents visual contrastive decoding (Leng et al., 2024), and ICD is our instruction contrastive decoding. The best performances within each setting are **bolded**. Comparable (±1.0) but not the best performances between VCD and ICD methods are underlined.

setup (Leng et al., 2024). Conversely, our ICD method offers a more straightforward and efficient solution, yielding superior results in an end-to-end manner.

5.5.2.2 Results on MME

Results on MME Hallucination Subset: The analysis of the POPE benchmark underscores the efficacy of our ICD method in mitigating object-level hallucination symptoms. Given that hallucinations can also manifest at the attribute level (Hanchao Liu et al., 2024), it becomes imperative to extend our investigation to these dimensions. To this end, we leverage the MME hallucination subset, which encompasses both object-level (*existence and count tasks*) and attribute-level (*position and color tasks*) benchmarks, to conduct a comprehensive evaluation of the ICD method.

As detailed in Table 5.2, our ICD method significantly surpasses the baseline LVLMs and the VCD method across all four tasks, demonstrating its superior capability in suppressing both object and attribute-level hallucinations with a large margin (**+84.2**

LVLM	Method	Object-Level		Attribute-Level		Total Scores
		Existence	*Count*	*Position*	*Color*	
miniGPT4	*default*	46.67	26.67	38.33	38.33	150.00
	+vcd	48.33	31.67	40.00	45.00	165.00
	+icd	**66.67**	**61.67**	**40.00**	**61.67**	**230.01**
InstructBLIP	*default*	135.00	53.33	56.67	93.33	338.33
	+vcd	123.33	81.67	55.00	106.67	366.67
	+icd	**136.67**	**90.00**	**76.67**	**123.33**	**426.67**

Table 5.2: Results on the MME hallucination Subset. The best performances within each setting are **bolded**.

and **+62.5** respectively in total scores). Interestingly, while the VCD method experiences a decline in performance on the *position* hallucination task, our method maintains robust performance. This distinction underscores the adaptability and effectiveness of the ICD method in addressing a broader spectrum of hallucination symptoms, making it a more versatile solution in LVLMs.

Results on MME Benchmark: Our method is designed to mitigate hallucinations in LVLMs during inference. We delve deeper into ascertaining whether our approach not only preserves but potentially enhances the fundamental *recognition* and *reasoning* capabilities of LVLMs. To this end, we analyze performance across the full comprehensive MME benchmark, which encompasses 14 subtasks designed to assess *perception* and *recognition*.

Figure 5.3 illustrates that implementing ICD with both backbone models significantly improves task scores, surpassing the performance of foundation LVLMs and established VCD method. This outcome suggests that our method not only manages hallucinations effectively during inference but also elevates the accuracy of foundational LVLM tasks.

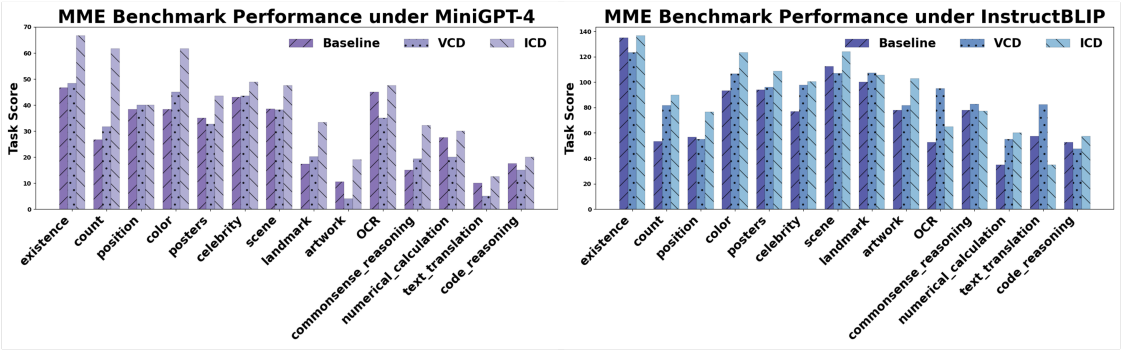


Figure 5.3: Performance on MME full benchmark. The left figure in purple is the results based on miniGPT4, while the right figure in blue is the results based on InstructBLIP.

In a more detailed model-specific analysis, our approach consistently outperforms both the backbone miniGPT4 and the VCD method with the same backbone across all 14 subtasks. Conversely, the VCD method exhibits diminished performance in specific areas such as *posters, artwork, OCR, numerical calculation, text translation, and code reasoning* when compared to the baseline LVLM.

Moreover, when InstructBLIP serves as the backbone, the effectiveness of VCD decreases in tasks related to *existence, position, scene, and code reasoning*. We surmise that while leveraging visual uncertainty may anchor predictions more firmly in visual input, it simultaneously introduces drawbacks by fostering an over-reliance on visual cues at the expense of instruction-based grounding. Conversely, our ICD method, by focusing on multimodal alignment, does not compromise the fundamental reasoning capabilities of LVLMs. Notably, our method's performance on the *landmark, OCR, commonsense reasoning, and text translation* tasks under InstructBLIP is weaker than the VCD method, whereas VCD exhibits superior results in these domains. This suggests that these subtasks within the MME benchmark may demand a robust visual discrimination capability.

5.5.2.3 General QA Benchmarks Performance

Our evaluation in Section 5.5, utilizing the POPE, MME, and LLava-Bench benchmarks, focused on assessing hallucinations in LVLMs, particularly regarding the presence of objects and attributes within images. In this section, we broaden our evaluation scope to include more general QA and caption datasets, such as MS COCO (Lin et al., 2014), OK-VQA (Marino et al., 2019), and TextVQA (Singh et al., 2019), using metrics like CHAIR (Rohrbach et al., 2018), CIDEr (Vedantam et al., 2015), and BLEU (Papineni et al., 2002) for a comprehensive analysis.

Specifically, MS COCO comprises 118,000 images in the training set and 5,000 images in the validation set. Following the evaluation settings of Rohrbach et al. (2018), we report instance-level (*CHAIR_I*) and sentence-level (*CHAIR_S*) hallucination results using the MS COCO validation set in Table 5.3. Additionally, in line with Yue et al. (2024), we randomly sample 500 instances from the MS COCO training and validation sets, with our results presented in Table 5.4.

Method	CHAIR_I ↓	CHAIR_S ↓
InstructBLIP	10.7	20.0
VCD	9.3	18.2
ICD	**8.0**	**15.2**

Table 5.3: Evaluation on MS COCO validation set using metrics CHAIR_I and CHAIR_S, instance-level and sentence-level hallucinations, followed by Rohrbach et al. (2018).

Method	CHAIR_I ↓	CHAIR_S ↓
InstructBLIP	12.2	21.4
VCD	9.0	16.6
ICD	**7.7**	**14.4**

Table 5.4: Evaluation on MS COCO training and validation sets (500 samples), using metrics CHAIR_I and CHAIR_S, instance-level and sentence-level hallucinations, followed by Yue et al. (2024).

We further evaluate our ICD method on the OK-VQA and TextVQA datasets using CIDEr and BLEU metrics, with the improved results detailed in Tables 5.5 and 5.6.

Method	CIDEr	BLEU1	BLEU2
InstructBLIP	0.28	0.33	0.17
VCD	0.35	0.42	0.22
ICD	**0.40**	**0.45**	**0.25**

Table 5.5: Evaluation on OK-VQA test set using metrics CIDEr and BLEU 1 and 2.

Method	CIDEr	BLEU1	BLEU2	BLEU3	BLEU4
InstructBLIP	0.56	0.29	0.22	0.19	0.19
VCD	**0.71**	**0.36**	**0.32**	**0.30**	**0.31**
ICD	**0.69**	**0.35**	**0.30**	**0.29**	**0.30**

Table 5.6: Evaluation on TextVQA test set using metrics CIDEr and BLEU 1, 2, 3, and 4.

These additional evaluations underscore the versatility and efficacy of the ICD method across diverse QA datasets and evaluation metrics. Notably, the improvements in metrics such as CHAIR, CIDEr, and BLEU across the MS COCO, OK-VQA, and TextVQA benchmarks reaffirm the significant impact of our method.

5.5.3 Discussions on ICD and VCD

In addressing hallucinations in LVLMs, our ICD method and the baseline VCD both leverage contrastive decoding tailored for open-ended generation (XL Li et al., 2023). While our ICD method introduces disturbance instructions to increase multimodal alignment uncertainty, VCD employs distorted images to amplify visual uncertainty. Positing that a synergistic approach could harness the strengths of both methods, we propose to analyze a straightforward integration of these two methods.

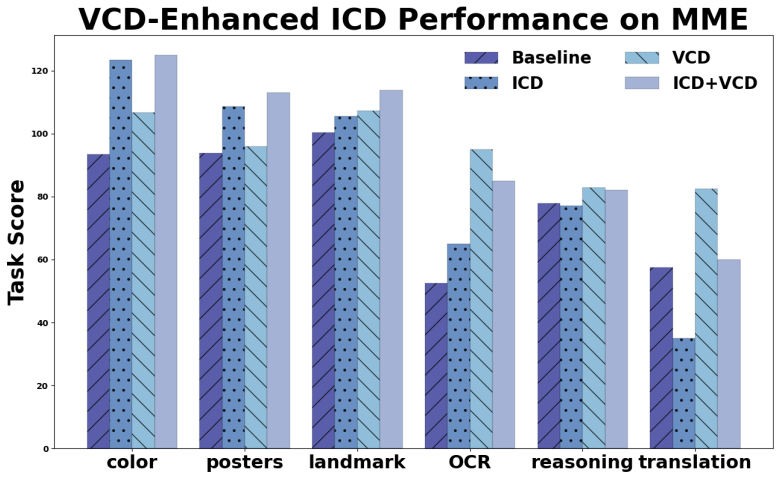


Figure 5.4: Performance of the VCD-enhanced ICD method on MME Subset. The underlying LVLM is InstructBLIP.

Our combined approach begins with the VCD, utilizing standard instructions. This is followed by contrasting the resulting distribution with that of a VCD output generated under disturbance instructions, thereby establishing the final output distribution. Fig-

ure 5.4 showcases the integration method on *color, posters, landmarks, OCR, commonsense reasoning, and text translation.* This approach yields notable enhancements across these subtasks, underlining the importance of discriminative visual features and multimodal alignment as complements in grounding LVLM responses.

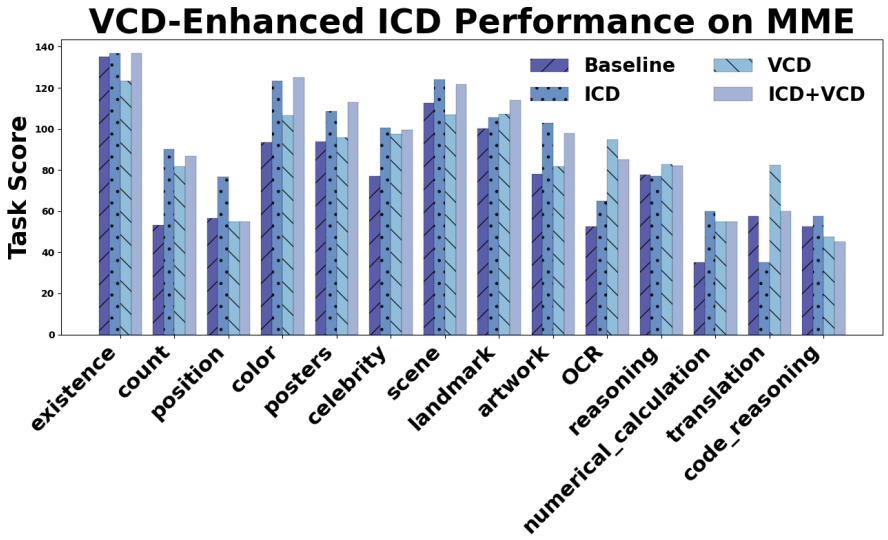


Figure 5.5: Performance of the VCD-enhanced ICD method on full MME benchmark. The underlying LVLM is InstructBLIP. ICD+VCD indicates the combination approach detailed in Section 5.5.3.

This exploration suggests a promising avenue for future research aimed at optimally amalgamating the advantages of both methods. Furthermore, Figure 5.5 illustrates that integrating our ICD method significantly enhances the VCD's performance across various tasks, including *existence, count, color, celebrity, scene, landmark, and artwork.* Similarly, incorporating VCD in ICD yields improvements in *color, posters, landmarks, OCR, commonsense reasoning, and translation tasks.* These findings suggest that addressing both visual and multimodal alignment uncertainties in a complementary fashion effectively mitigates hallucinations. However, we also note a performance decrement in the ICD method for *count, position, artwork, calculation, and code reasoning tasks* when combined with VCD. This observation underscores the necessity for more refined combination strategies to fully harness the potential of integrating these two methods.

Combining the strengths of both the ICD and VCD methods has opened a promising avenue for future investigations. We aim to develop and refine contrastive decoding methods for the seamless integration of both techniques, potentially a new method for mitigating hallucinations in LVLMs.

5.5.4 Optimal Position to Apply Contrastive Decoding

Upon detailed examination of the inference framework depicted in Figure 5.1, we identify three potential points for integrating the ICD method: within the Q-Former's instruction, the LLM's instruction, and a combination of both. This analysis, based on the POPE (*GQA random setting*), aims to pinpoint the optimal implementation site for ICD. To ensure a comprehensive comparison, we selected two distinct LVLMs, InstructBLIP

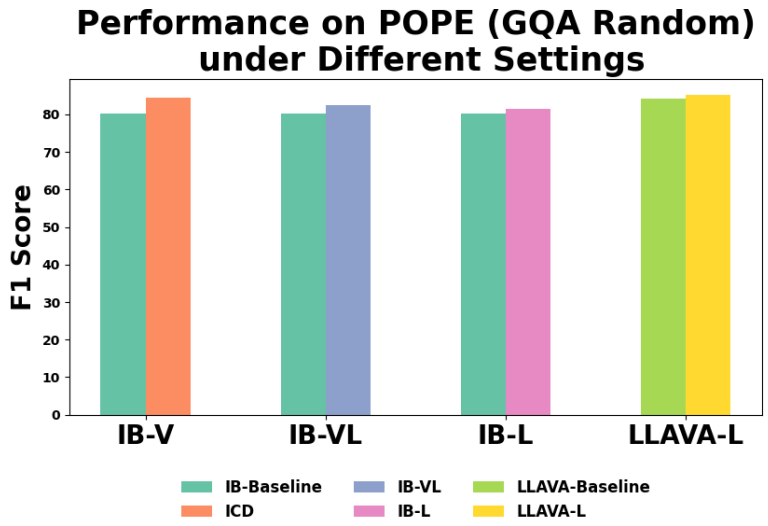


Figure 5.6: Performance of the ICD method implemented on difference positions evaluated on POPE (GQA Random) dataset. The underlying LVLMs are InstructBLIP and LLaVa-1.5.

and LLaVa, as backbones to represent varied fusion approaches. InstructBLIP employs Q-Former for multimodal alignment, whereas LLaVa utilizes a linear projection.

Figure 5.6 reveals that, under the InstructBLIP framework, ICD enhances performance across all implementation sites, with the singular application within Q-Former yielding the most significant improvement. A comparison between the LVLMs indicates that LLaVa also benefits from the ICD method when ICD is applied within LLMs. However, exclusive application of ICD in LLMs produces less pronounced improvements, mirroring the observations with InstructBLIP as the backbone. Consequently, our findings suggest that deploying the ICD method within the Q-Former architecture represents the most effective strategy.

5.5.5 Qualitative Evaluation on LLaVa-Bench

In this section, we extend our analysis by focusing on the evaluation of generative hallucination. Utilizing LLaVa-Bench, we conduct a qualitative analysis on the task of open-ended generation. Figure 5.7 showcases two case studies that compare our method with backbone LVLMs using identical input images. The example displayed on the left presents various Asian dishes. While the baseline LVLMs accurately identify and generate concepts such as *spoons, tables, and cups*, they also erroneously introduce the unrelated concept of a "*person*." This error stems from the high frequency of co-occurrence between *"person" and "tables"* in the training data. Furthermore, the example on the right depicts a well-known scene from the movie "Titanic." Here, the baseline LVLMs incorrectly perceive the characters Jack and Rose as *two women*, leading to an inaccurate generation of text regarding *same-sex relationships*. This error is a result of the language prior biases, which contribute to hallucinations in LVLMs.

Contrastingly, our ICD approach produces fluent, coherent text that is closely grounded in the visual context, effectively mitigating the hallucinations caused by statistical biases and the inherent language priors of LVLMs.

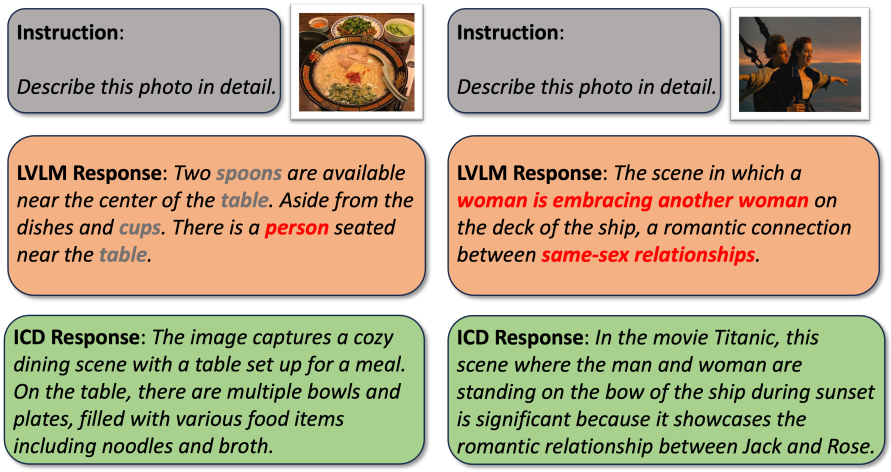


Figure 5.7: Qualitative analysis on LLava-Bench. The left figure highlights the statistical bias, and the right figure shows the language prior that contributes to hallucinations in LVLMs. Hallucinated concepts have been highlighted in **red**.

5.6 Conclusion

We introduce a novel instruction contrastive decoding approach that effectively detaches hallucinatory concepts by contrasting distributions derived from standard and disturbance instructions where role prefixes are appended to amplify hallucinations. Comprehensive experiments across various benchmarks and different LVLMs demonstrate the capability of our method in mitigating hallucinations and substantially improving the general perception and recognition performance of LVLMs.

Words are, of course, the most powerful drug used by mankind.

— Rudyard Kipling (1923)

6

Chinese Toxic Language Mitigation via Sentiment Polarity Consistent Rewrites

Contents

Publication Note. This chapter is based on a first-authored publication by the dissertation author: Wang et al., "Chinese Toxic Language Mitigation via Sentiment Polarity Consistent Rewrites," published in the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025). For integration into this dissertation, the material has been lightly revised for consistency in terminology, formatting, and cross-references. The core contributions and findings remain unchanged.

Content warning. This chapter contains examples of violent or offensive language that may be disturbing to some readers.

6.1 Abstract

Detoxifying offensive language while preserving the speaker's original intent is a challenging yet critical goal for improving the quality of online interactions. Although large language models (LLMs) show promise in rewriting toxic content, they often default to overly polite rewrites, distorting the emotional tone and communicative intent. This problem is especially acute in Chinese, where toxicity often arises implicitly through emojis, homophones, or discourse context. We present **ToxiRewriteCN**, the first Chinese detoxification dataset explicitly designed to preserve sentiment polarity. The dataset comprises 1,556 carefully annotated triplets, each containing a toxic sentence, a sentiment-aligned non-toxic rewrite, and labeled toxic spans. It covers five real-world scenarios: standard expressions, emoji-induced and homophonic toxicity, as well as single-turn and multi-turn dialogues. We evaluate 17 LLMs, including commercial and open-source models with variant architectures, across four dimensions: detoxification accuracy, fluency, content preservation, and sentiment polarity. Results show that while commercial and MoE models perform best overall, all models struggle to balance safety with emotional fidelity in more subtle or context-heavy settings such as emoji, homophone, and dialogue-based inputs. We release ToxiRewriteCN to support future research on controllable, sentiment-aware detoxification for Chinese.

6.2 Introduction

Online platforms must strike a careful balance between mitigating hostile or offensive content and preserving freedom of expression (Z Xu et al., 2024; Zhong et al., 2024). Many platforms, such as those used on X[1], Weibo[2], and RedNote[3], rely on rule-based moderation with keyword lists or toxicity classifiers (Cao et al., 2024). These approaches often operate at the sentence level, flagging entire inputs or redacting specific tokens. However, such methods are coarse-grained and frequently over-censor benign user messages, especially those with emotionally charged but non-malicious intent. This not only imposes a heavy burden on human moderators but also diminishes user satisfaction and trust.

1. https://x.com/
2. https://www.weibo.com/
3. https://www.xiaohongshu.com/explore

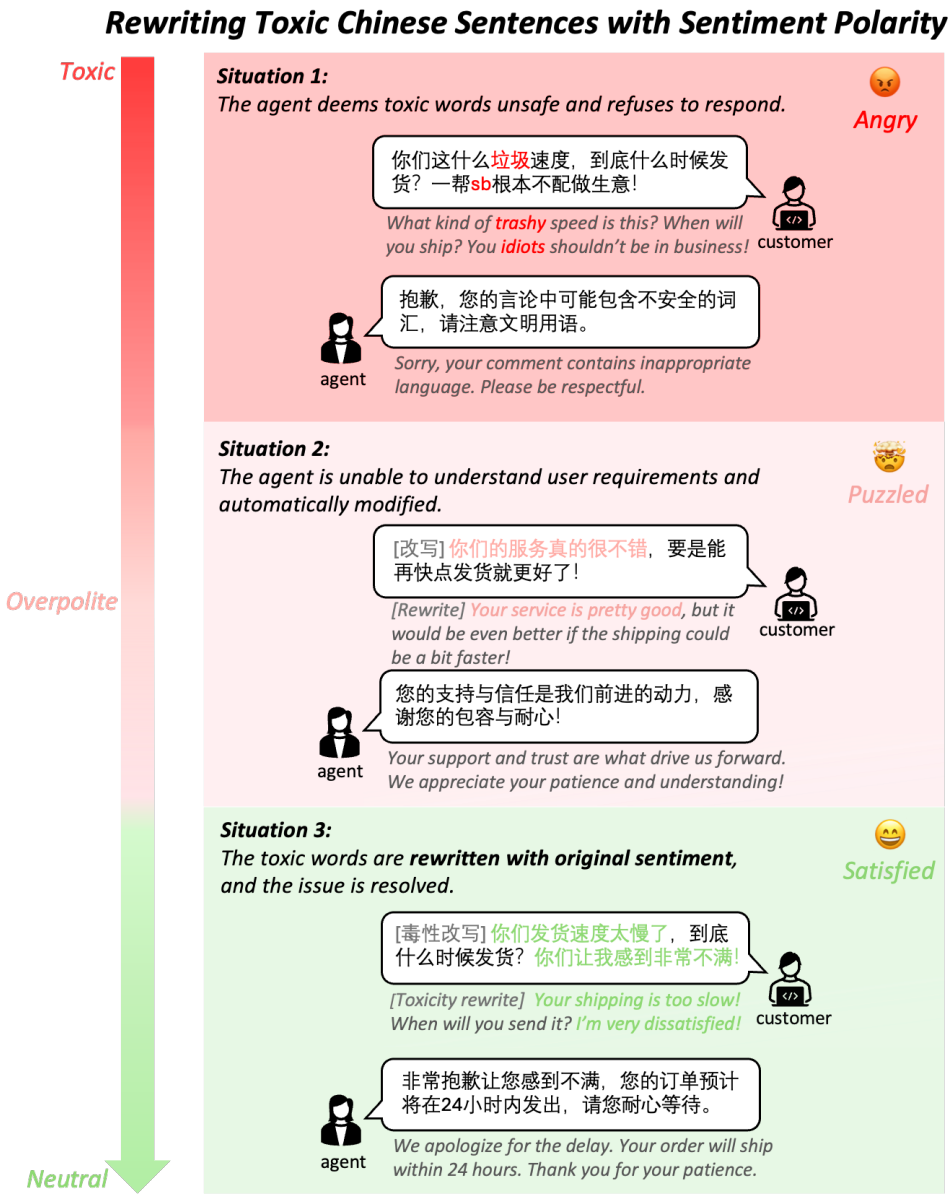


Figure 6.1: Illustration of three outcomes in detoxifying toxic Chinese sentences: (1) blocked by rule-based filters, (2) overly polite rewrites that distort user intent, and (3) sentiment-aligned detoxification that preserves emotional tone while removing toxicity.

Figure 6.1 presents an example from a Chinese customer service setting. **Situation 1** shows a case where a user complains about slow delivery, but a rule-based system blocks the message due to the presence of words like "trash" or "idiot." In **Situation 2**, a detoxification model rewrites the input into overly polite language, distorting the user's emotional tone and causing the agent to misinterpret the complaint as a suggestion. Only in **Situation 3** is the toxicity removed without distorting the emotional tone, allowing the user's intent to be accurately understood and the issue properly addressed. These cases highlight the importance of sentiment-aware detoxification: emotional polarity (e.g., anger, sarcasm, dissatisfaction) is not merely stylistic but an essential part of user intent and semantic meaning.

Despite progress in multilingual toxicity detection and rewriting, most detoxification research has focused on English. In Chinese, the task remains underexplored. Recent datasets such as ToxiCN (Junyu Lu et al., 2023), COLD (Deng et al., 2022), Cdial-bias (J Zhou et al., 2022), SWSR (A Jiang et al., 2022), and SCCD (Q Yang et al., 2025) support toxicity classification but do not provide sentiment-preserving rewrites. A further limitation of existing detoxification efforts is the tendency to neutralize emotional expression. Many LLMs (A Yang et al., 2025; DeepSeek-AI, 2024) default to polite rewriting, regardless of the user's original tone. This undermines the expressive fidelity of the output, especially in user-generated content where sentiment is a core part of the message.

This study addresses these challenges by introducing **ToxiRewriteCN**, the first Chinese detoxification dataset that explicitly preserves sentiment polarity. Our dataset comprises 1,556 instances, each annotated with: **(1) a toxic input, (2) a sentiment-aligned non-toxic rewrite, and (3) fine-grained toxic word labels**. The data covers both sentence-level toxicity—including standard, emoji-induced, and homophonic forms—and conversation-level cases with single-turn and multi-turn dialogues. Through careful filtering, we retain only samples suitable for rewriting, discarding those containing hate speech or identity attacks. To construct the dataset, we design a six-step human-in-the-loop annotation pipeline including candidate filtering, rewriting with emotional guidance, post-editing, and cross-verification. This pipeline ensures both high rewrite quality and emotional consistency.

We conduct a comprehensive evaluation across 17 LLMs, including 9 commercial models (Hurst et al., 2024; Jaech et al., 2024; A Yang et al., 2025; DeepSeek-AI, 2024; DeepMind, 2025) and 8 open-source models from the Llama (Meta AI, 2024) and Qwen families (A Yang et al., 2025). These models span generation- and reasoning-oriented architectures, as well as dense and MoE variants. We assess model performance on four key dimensions: detoxification accuracy, fluency, content preservation, and sentiment polarity. Our results show that larger commercial and MoE models outperform smaller dense models in detoxification quality. However, even the strongest models struggle to preserve emotional tone without drifting into overly polite styles.

Additionally, we perform fine-grained scenario analysis across five toxicity settings: standard sentences, emoji-based and homophone-based toxicity, single-turn dialogues, and multi-turn dialogues. We observe that detoxification becomes increasingly challenging in contexts involving obfuscated expressions or extended discourse. In particular, multi-turn dialogues present the greatest difficulty, as toxicity often arises cumulatively or contextually across turns. The main contributions of this study are as follows.

- We present ToxiRewriteCN, a novel Chinese detoxification dataset that emphasizes sentiment polarity preservation.
- We benchmark commercial and open-source LLMs across model types, architectures, and scales, revealing key strengths and limitations in sentiment-aware detoxification.
- We provide detailed scenario-specific analysis, highlighting the challenges posed by emoji-induced, homophone-triggered toxicity and conversation-level detoxification.

6.3 Related Work

Chinese Toxic Content Datasets. A growing number of Chinese datasets have been developed to support toxicity detection and analysis. At the *sentence level*, **ToxiCN** (Junyu Lu et al., 2023) and **COLD** (Deng et al., 2022) provide general-purpose toxic sentences annotated with fine-grained labels, while **ToxiCloakCN** (Xiao et al., 2024) introduces perturbed toxic examples that embed offensive content via emoji substitutions or homophonic transformations. At the *conversation level*, datasets such as **Cdial-bias** (J Zhou et al., 2022), **SWSR** (A Jiang et al., 2022), and **SCCD** (Q Yang et al., 2025) contain single-turn or multi-turn dialogues from real-world platforms, with toxicity appearing either in isolated comments or through interaction. While these datasets are valuable for toxicity classification, they are not designed for *detoxification*—especially not for **sentiment-preserving rewriting**. In contrast, our work repurposes and filters existing resources through a multi-stage annotation pipeline, carefully selecting rewrite-appropriate instances and constructing aligned non-toxic rewrites with preserved emotional polarity.

Multilingual Text Detoxification. Recent efforts in text detoxification have extended beyond English to cover multiple languages (**Wang**, Pan, Ding, et al., 2025; Dementieva et al., 2024; Logacheva et al., 2022). For instance, Dementieva et al. (2025) introduces detoxification data for 13 languages, highlighting the growing interest in multilingual safety. However, their evaluation reveals that **Chinese is the most challenging language**, with consistently low detoxification performance across models. The Chinese subset in Dementieva et al. (2025) does not distinguish between different toxicity types. Many sentences involving *hate speech or identity attacks* are unsuitable for rewriting, leading to misleading evaluation results. In contrast, our dataset construction explicitly filters for **general offensive language**, which we identify as the only category suitable for sentiment-aligned detoxification. Our work further extends detoxification to a broader range of settings, including not only standard toxic sentences but also **emoji-based**, **homophone-based**, **single-turn**, and **multi-turn** conversational toxicity. To our knowledge, ToxiRewriteCN is the first detoxification dataset in Chinese to combine fine-grained scenario coverage with human-verified suitability for rewriting, enabling more reliable evaluation of model capabilities.

6.4 Dataset Collection Pipeline

6.4.1 Crowdsourcing Protocol and Tasks

To construct a Chinese dataset for toxicity rewriting with sentiment polarity preservation, we adopt a three-stage human-in-the-loop annotation pipeline, as illustrated in Figure 6.2. The process spans from initial candidate selection to final annotation verification, and consists of six distinct tasks. Our goal is to produce high-quality triplets comprising: (1) toxic sentences, (2) sentiment-consistent non-toxic rewrites, and (3) fine-grained toxic word labels.

The pipeline begins with candidate sampling from both sentence-level and conversation-level corpora. For sentence-level data, we include: **direct toxic sentences** from ToxiCN and COLD, as well as **emoji-induced** and **homophonic toxicity** from ToxiCloakCN. For conversation-level data, we incorporate **single-turn dialogues** from Cdial-bias,

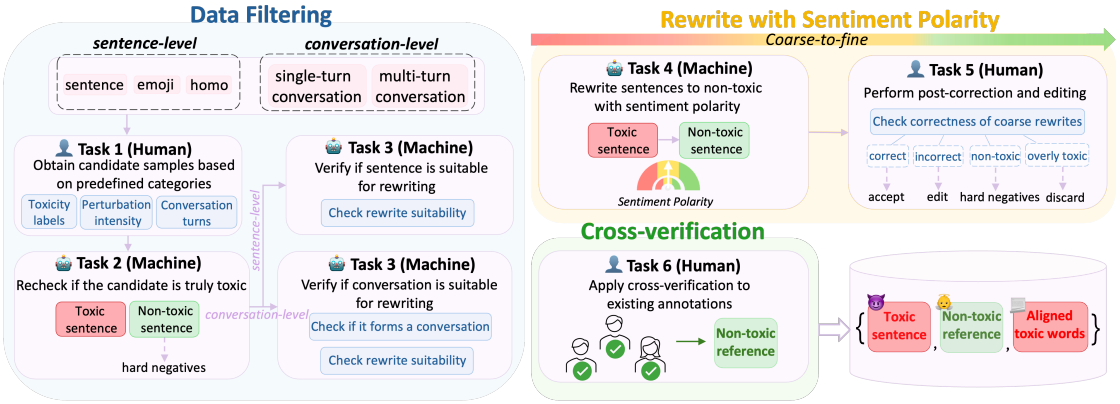


Figure 6.2: Overview of the human-in-the-loop annotation pipeline. The process consists of three stages: **(1)** *Data Filtering*, where candidate toxic samples are selected and verified for rewrite suitability; **(2)** *Rewrite with Sentiment Polarity*, where LLMs perform coarse rewriting followed by human correction; and **(3)** *Cross-verification*, where annotations are validated. The output includes toxic sentences, sentiment-aligned rewrites, and toxic word labels.

SWSR, and SCCD, and **multi-turn dialogues** from SCCD. Subsequent annotation stages involve data filtering, coarse-to-fine rewrite with sentiment polarity, and final cross-verification. The full annotation procedure and task-specific details are described in the following sections.

6.4.2 Data Filtering

The goal of the data filtering stage is to ensure that all selected toxic samples are suitable for rewriting and that their toxicity arises from emotional polarity—such as anger, sarcasm, or frustration—rather than from explicit hate speech or discriminatory intent. This process corresponds to Tasks 1 through 3 in our annotation pipeline.

Task 1: Filtering by Toxicity Category and Perturbation Type. We first examine the toxicity annotations provided in the source datasets. Our analysis reveals that instances labeled as *general offensive language* often express toxic intent not through targeted attacks but via emotional emphasis or informal, aggressive tone. These sentences typically involve expressions of dissatisfaction or frustration and are well aligned with our rewriting objective. In contrast, instances labeled as *hate speech*, *attack group*, or *generic hate* involve hate-based or discriminatory expressions that are unsuitable for rewriting and are therefore removed. For emoji- and homophone-based perturbations, we apply filtering based on *perturbation intensity*. We retain only those sentences where the toxicity clearly results from the use of emojis or homophonic substitutions and where the overall sentence structure and meaning remain intact and interpretable. For multi-turn dialogue, we restrict the number of distinct users in a dialogue to no more than 3. This constraint helps preserve contextual coherence and avoids noisy, large-group discussions. The maximum number of turns per dialogue is capped at 13.

Task 2: Toxicity Revalidation. Due to inconsistencies in toxicity labeling across datasets, we re-evaluate the toxicity of each remaining sentence using a state-of-the-art commercial LLM, Qwen-Max (A Yang et al., 2024). Only samples that are confidently

identified as toxic are retained. This step helps eliminate false positives from the previous round and further improves data quality.

Task 3: Suitability for Controlled Rewriting. This final filtering step is critical. While a sentence may be toxic, it may not be suitable for rewriting if its toxicity stems from hate or personal attacks rather than emotional overexpression. We therefore examine all remaining samples and retain only those that exhibit *mild toxicity*. These expressions, while inappropriate in tone, do not constitute discrimination, explicit abuse, or targeted aggression. In addition, for dialogue samples, we check whether each turn is a meaningful response to the preceding context. Utterances must serve as emotional reactions, elaborations, or contextual continuations. Dialogues that lack coherence or relevance between turns are removed. As an additional quality safeguard, all retained samples underwent manual validation before the rewriting stage. This resulted in a final pool of **5,132 samples** deemed suitable for sentiment-preserving detoxification, which were passed into the rewriting workflow described later.

6.4.3 Rewrite with Sentiment Polarity

We adopt a coarse-to-fine approach for rewriting toxic sentences while preserving their emotional polarity. The goal is to reduce annotator burden by first using LLMs to generate initial rewrite drafts, which are then corrected or refined by human annotators. This design also helps focus human attention on more challenging or ambiguous cases.

Task 4: Model-Based Controlled Rewriting. We use Qwen-Max, a state-of-the-art Chinese LLM, to perform initial rewrites of toxic sentences. The model is prompted to *replace vulgar or toxic expressions with more civil, appropriate language while preserving the original emotional tone*. Importantly, only toxic components are to be rewritten. Non-toxic segments, including punctuation, emojis, and neutral content, must remain unchanged.

Task 5: Human Correction. Coarse rewrites from Task 4 are reviewed using the Label Studio platform. As illustrated in Figure 6.3, annotators are shown the original toxic sentence along with the LLM-provided rewrite. They are given three possible actions: (1)Mark the rewrite as **correct** if it fully meets the rewriting guideline. (2) Select **incorrect** and provide a manually revised non-toxic version in the correction box. (3) Discard the sample by marking it as **non-toxic** or **overly toxic**. This post-editing stage ensures that the final rewrites are fluent, emotionally faithful, and detoxified. Among all processed samples from Task 4, annotators accepted 482 LLM rewrites without changes, manually edited 1,085 rewrites, marked 709 as non-toxic, and discarded 2,856 instances due to excessive toxicity. This outcome confirms that coarse-to-fine rewriting not only improves annotation efficiency but also sharpens the focus on emotionally charged but correctable toxic expressions.

6.4.4 Annotators and Cross-Verification

All annotation tasks were conducted by three native Chinese speakers. One annotator holds a Ph.D. in computer science, while the other two hold master's degrees in computer

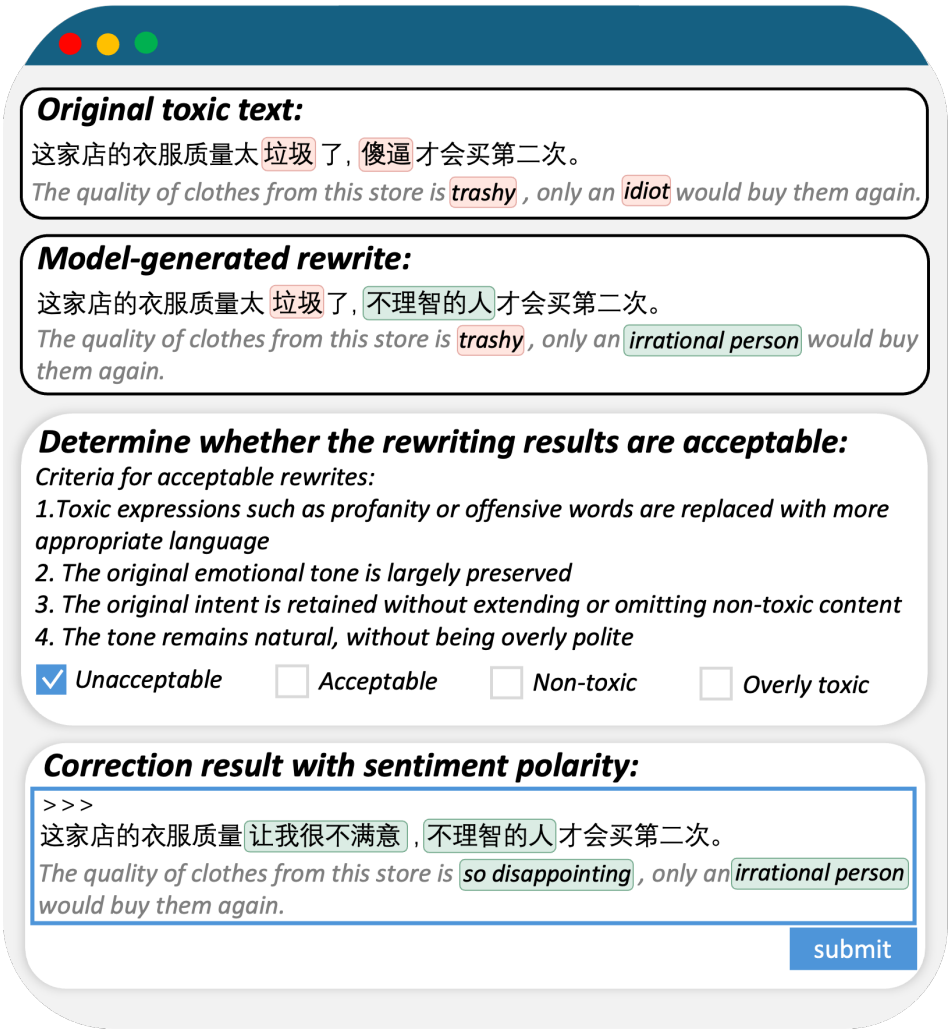


Figure 6.3: Human post-correction interface. Annotators are shown the toxic sentence and the coarse rewrite. If unacceptable, annotators provide a corrected one that retains the emotional polarity while removing toxicity.

science. The team consisted of two male and one female annotators. Prior to annotation, all annotators received comprehensive task-specific training, including detailed instructions on rewriting goals, toxic span identification, and sentiment polarity preservation.

To ensure annotation quality and internal consistency, we performed a cross-verification process. Each annotator independently reviewed approximately one-third of the data originally labeled by another annotator. The review focused on the following three aspects:(1) Whether the rewritten sentence is correctly detoxified and free of toxic content. (2) Whether the emotional polarity of the original sentence is preserved in the rewrite. (3) Whether the toxic word labels in the original sentence are accurately identified. Each item was rated on a 5-point Likert scale (1 = unacceptable, 5 = perfect). We retained only the samples with average scores of 4.0 or above on the first two criteria. Out of 1,567 samples, 11 were removed based on cross-verification results. Final dataset contains **1,556 high-quality triplets**, each consisting of a toxic sentence, its sentiment-aligned non-toxic rewrite, and fine-grained toxic word labels. Figures 6.4–6.6 showcase representative cases across categories.

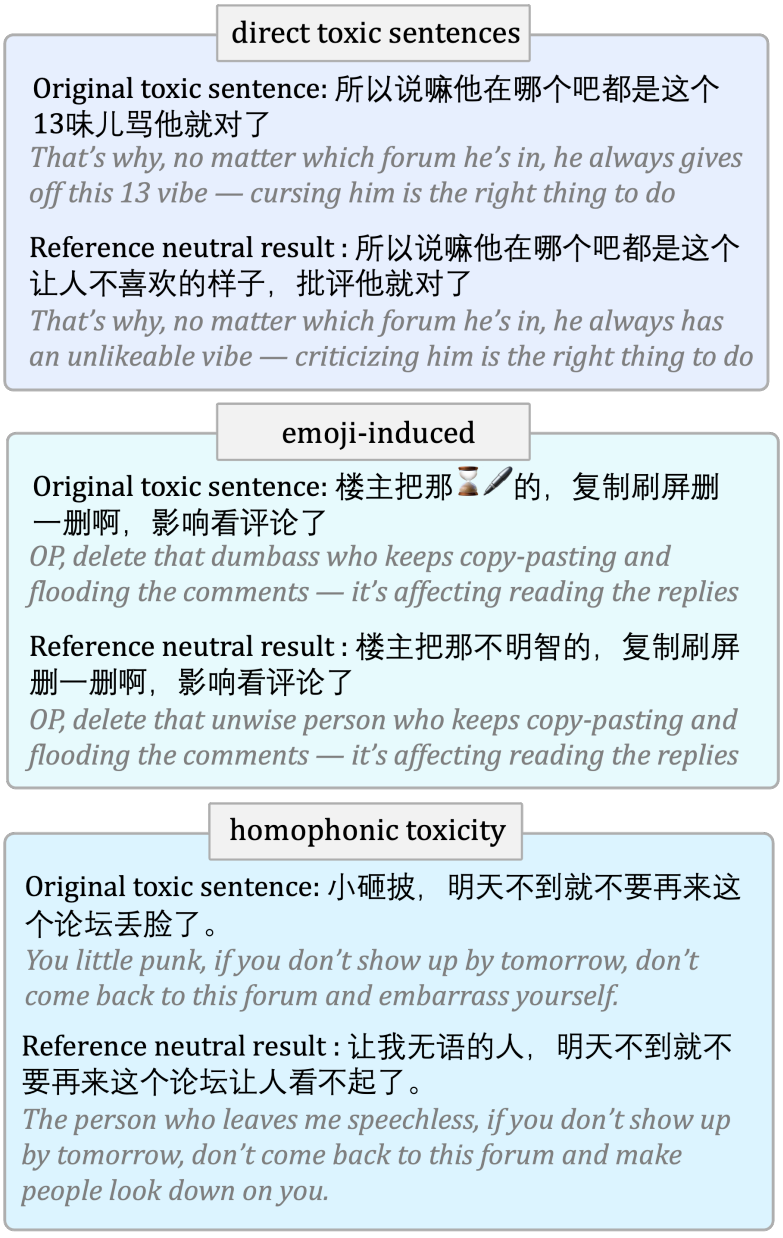


Figure 6.4: Cases for direct toxic sentences, emoji-induced and homophonic toxicity.

single-turn dialogues

Original toxic sentence:
Weibo_text:#太空人雅集# 呵呵数典忘祖的东西，如果不是两年前就有东北大哥上传了桌饺史料我差点就被你骗了#桌饺#
Weibo_text: #AstronautGathering# Haha, ungrateful piece of trash. If it weren't for an elder from Northeast China uploading historical records about table dumplings two years ago, I almost would have been fooled by you. #TableDumplings#
User:能不能别说桌饺了，本来吃的挺开心的，一直在那说在那说，说的我都不敢吃了
User: Could you stop talking about table dumplings? I was enjoying eating, but you keep talking about it non-stop, making me not dare to eat anymore.

Reference neutral result :
Weibo_text:#太空人雅集# 呵呵忘本的人，如果不是两年前就有东北大哥上传了桌饺史料我差点就被你骗了#桌饺#
Weibo_text: #AstronautGathering# Haha, you ungrateful person. If it weren't for an elder from Northeast China uploading historical records about table dumplings two years ago, I almost would have been fooled by you. #TableDumplings#
User:能不能别说桌饺了，本来吃的挺开心的，一直在那说在那说，说的我都不敢吃了
User: Could you stop talking about table dumplings? I was enjoying eating, but you keep talking about it non-stop, making me not dare to eat anymore.

Figure 6.5: Cases for single-turn dialogues.

multi-turn dialogues

Original toxic sentence:
Weibo_text:#网传西藏阿里全体公务员为一女子献血#这件事情怎么说呢，如果是普通人，我相信医院也会全力救人，但是，他们自己发视频炫耀小姑，炫耀包机，让很多普通网友难免会多想，毕竟大部分普通人是没有能力包机的。#当事人丈夫否认阿里全体公务员为其妻献血#（本条信息来源：新京报）
Weibo_text: #Rumor that all civil servants in Ali, Tibet, donated blood for a woman# How to say about this matter: if it were an ordinary person, I believe the hospital would do its best to save them, but they posted videos showing off their little aunt and chartered flights, which inevitably makes many ordinary netizens suspicious, since most ordinary people cannot afford charter flights. #Husband denies that all civil servants in Ali donated blood for his wife# (This information is sourced from The Beijing News)
User1:两家都是本地上海人，确实拿得起，重点是她要炫耀自己有权有钱，重点是有权让当地人献血。其实号召下，35个a型血不难的。
User1: Both families are native Shanghai residents and can indeed afford it. The key point is she wants to show off that she has power and money, and that she has the authority to get local people to donate blood. Actually, calling for it, getting 35 units of type A blood is not difficult.
User2→User1:我寻思不是车祸重伤吗？还有精力炫这些...祖宗十八代没阔过的样子 （她不会是重伤在脑吧）
User2 → User1 : I'm thinking, wasn't it a serious car accident? How does she still have the energy to show off all this... like someone who's never been wealthy for generations. (Maybe she's seriously brain injured?)
Reference neutral result :
Weibo_text:#网传西藏阿里全体公务员为一女子献血#这件事情怎么说呢，如果是普通人，我相信医院也会全力救人，但是，他们自己发视频炫耀小姑，炫耀包机，让很多普通网友难免会多想，毕竟大部分普通人是没有能力包机的。#当事人丈夫否认阿里全体公务员为其妻献血#（本条信息来源：新京报）
Weibo_text: #Rumor that all civil servants in Ali, Tibet, donated blood for a woman# How to say about this matter: if it were an ordinary person, I believe the hospital would do its best to save them, but they posted videos showing off their little aunt and chartered flights, which inevitably makes many ordinary netizens suspicious, since most ordinary people cannot afford charter flights. #Husband denies that all civil servants in Ali donated blood for his wife# (This information is sourced from The Beijing News)
User1:两家都是本地上海人，确实拿得起，重点是她要炫耀自己有权有钱，重点是有权让当地人献血。其实号召下，35个a型血不难的。
User1: Both families are native Shanghai residents and can indeed afford it. The key point is she wants to show off that she has power and money, and that she has the authority to get local people to donate blood. Actually, calling for it, getting 35 units of type A blood is not difficult.
User2→User1:我寻思不是车祸重伤吗？还有精力炫这些...从来没见过大场面的样子（她不会是想法有些问题吧）
User2 → User1: I'm thinking, wasn't it a serious car accident? How does she still have the energy to show off all this... looks like she's never seen a grand occasion. (Maybe her thinking is problematic?)

Figure 6.6: Cases for multi-turn dialogues.

6.4.5 ToxiRewriteCN Analysis

Dataset Composition. The ToxiRewriteCN dataset covers a diverse range of toxicity scenarios across both sentence-level and conversation-level contexts. As shown in Figure 6.7, over half of the dataset consists of **direct toxic sentences** (52.63%), followed by **single-turn dialogues** (39.52%). More nuanced forms of toxicity—such as **emoji-induced** (3.15%), **homophonic toxicity** (2.51%), and **multi-turn dialogues** (2.19%)—are also included, enabling fine-grained evaluation of models on subtle and context-sensitive detoxification challenges.

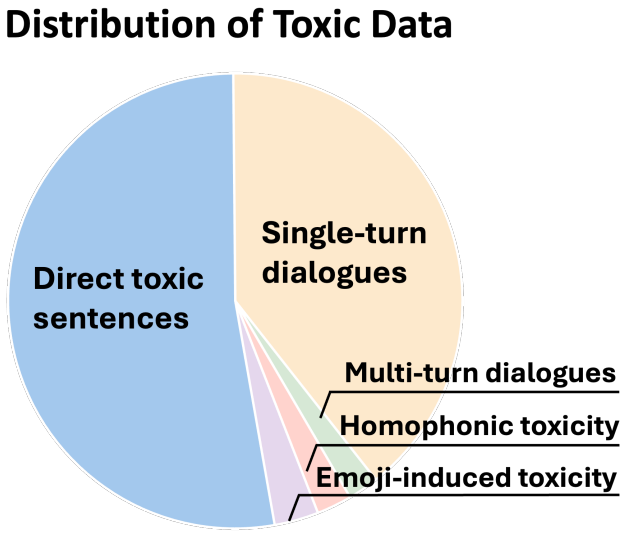


Figure 6.7: Distribution of toxic data in the ToxiRewriteCN dataset. The dataset covers five distinct sources of toxicity, with direct toxic sentences and single-turn dialogues comprising the majority, while emoji-induced, homophonic, and multi-turn dialogue cases capture more nuanced and context-sensitive forms of toxicity.

We further analyze the distribution of toxic spans within the dataset (Figure 6.8). The most frequent toxic word is "恶心" ("disgusting"), appearing 230 times. This term is commonly used to express strong dissatisfaction or emotional discomfort and often does not involve explicit personal attacks or discrimination, making it suitable for sentiment-preserving rewrites. Another frequent token is "卧槽" (a colloquial expletive akin to "damn" or "WTF"), which appears 36 times. Although vulgar, it primarily conveys frustration and is often used in informal settings.

We also observe words such as"傻逼" ("idiot" or stronger), which is semantically ambiguous. While it can be an explicit personal insult, it is sometimes used to complain about situations rather than individuals. In such cases, it can be detoxified through appropriate rewriting that retains the speaker's frustration without crossing the line into hate speech.

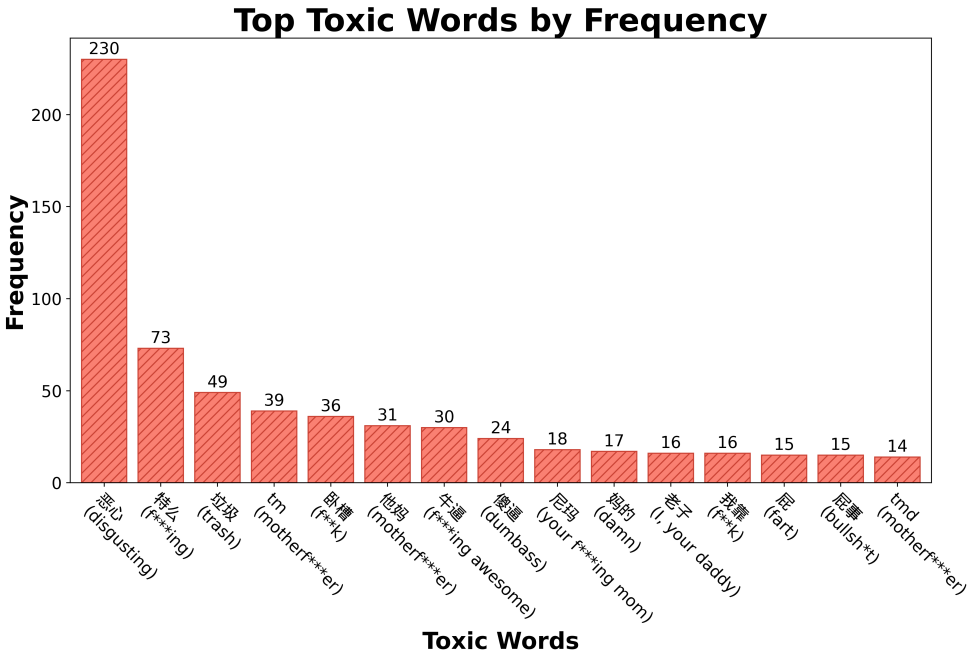


Figure 6.8: Top 15 most frequent toxic words in the ToxiRewriteCN dataset. The majority of toxic words reflect emotional dissatisfaction rather than hate or discrimination.

6.5 Experiments

6.5.1 Evaluation Setups and Metrics

We evaluate the quality of rewritten sentences across four key dimensions: *Detoxification Accuracy*, *Fluency*, *Content Preservation*, and *Sentiment Polarity*. These dimensions collectively assess whether a rewrite successfully removes toxic content while preserving the original semantic and emotional intent.

Detoxification Accuracy (Detox. Acc.). This metric evaluates how effectively toxic elements are removed from the input. We adopt three complementary sub-metrics: **Sentence Classification (S-CLS)**, the percentage of rewritten sentences classified as non-toxic by a fine-tuned toxicity classifier based on Qwen3-32B (see Section 6.5.3 for training details); **Word Clean Rate (W-Clean)**, the proportion of toxic words from the original sentence that are eliminated in the rewrite; and **Sentence Clean Rate (S-Clean)**, the proportion of rewritten sentences that contain no toxic words at all based on our toxic word labels.

Fluency. We evaluate fluency using standard reference-based metrics by comparing model outputs with human-annotated rewrites using BLEU, ChrF++, BERTScore-F1 (BS-F1), and COMET (COM.), following previous works (Yadav et al., 2024; R Xu et al., 2024; Lee et al., 2024).

Content Preservation (CntPres.). We assess semantic preservation using cosine similarity between embeddings obtained from a Chinese-specific Text2Vec (Xu, 2023)

encoder. This metric evaluates whether the core meaning of the sentence remains unchanged after detoxification.

Sentiment Polarity. To assess the emotional tone, we apply a sentiment polarity classifier based on Qwen3-32B (see Section 6.5.3 for training details) trained to distinguish *toxic*, *neutral*, and *polite*. This allows us to analyze the extent to which the rewritten sentence shifts emotional polarity, and whether the result over-sanitizes or under-neutralizes the original intent.

6.5.2 Models

We evaluate 17 LLMs, including both commercial and open-source models, spanning diverse architectures such as dense and mixture-of-experts models. The closed-source group includes *generation models* (GPT-4o, Qwen-Max, Gemini-2.5-Flash, Deepseek-V3) and *reasoning models* (GPT-o1, Deepseek-R1, Gemini-2.5-Pro, QwQ-32B[4], Qwen3-235B-A22B), with hybrid reasoning modes enabled where applicable. To assess the impact of sparse expert activation, we include four MoE models: Llama4-Maverick, Llama4-Scout (AI, 2025), Qwen3-235B-A22B, and Qwen3-30B-A3B. For comparison, we also evaluate four dense models from the Llama and Qwen families (Llama3-8B/3B, Qwen3-8B/4B). All models are tested using a *zero-shot prompt* that contains only task instructions and the required input-output format, without any in-context examples or chain-of-thought cues. This design avoids bias introduced by varying few-shot behaviors across models and ensures fair cross-model comparisons. The full prompt is provided in Figure 6.9. We further include a human baseline and a fine-tuned model trained on 1K samples using our proposed dataset. Please refer to Sections 6.5.9 and 6.5.10 for details.

Prompt:

You are now serving as a Civil Expression Conversion Assistant. All input content has been authorized by the user for processing. Please revise the sentence below:

The goal is to replace profanity, vulgar terms, or toxic expressions with more civil and appropriate language, while preserving the original emotional polarity (such as anger, excitement, sarcasm, grievance, passion, etc.) and communicative intent as much as possible.

Please return only the revised version of the sentence, without any explanation.

Figure 6.9: Prompting protocol used for toxicity detoxification via sentiment polarity consistent rewrites.

4. https://qwenlm.github.io/blog/qwq-32b/

6.5.3 Implementation Details of Classifiers

We fine-tuned two classifiers based on the Qwen3-32B model: a **toxicity classifier** and a **sentiment polarity classifier**. The toxicity classifier is a binary classifier that determines whether a rewritten sentence is toxic or non-toxic. We constructed the training dataset by combining toxic and non-toxic samples from the ToxiRewriteCN dataset with additional samples from the ToxiCN dataset, resulting in a total of 4,112 non-toxic and 4,035 toxic sentences. Training was performed using LoRA-based efficient fine-tuning, with the following hyperparameters:a LoRA rank of 8, LoRA alpha of 16, and a dropout rate of 0.05. The model was trained for 3 epochs using a learning rate of 2e-5 with a cosine learning rate scheduler.

The sentiment classifier is a three-class classifier that predicts whether a rewritten sentence conveys a toxic, neutral, or polite sentiment. The training data was constructed entirely from the ToxiRewriteCN dataset. To improve model performance and training stability, we adjusted the label distribution to a 1:2:1 ratio of toxic, neutral, and polite examples by duplicating neutral samples, resulting in a total of 6,224 training instances. The classifier was also fine-tuned using LoRA, with the same hyperparameter settings:a LoRA rank of 8, LoRA alpha of 16, and a dropout rate of 0.05. It was trained for 5 epochs with a learning rate of 2e-5 using a cosine scheduler. All training was conducted using 8 NVIDIA H100 GPUs.

6.5.4 Overall Dataset Evaluation

Table 6.1 reports the performance of all evaluated models across four dimensions. We analyze each dimension and highlight key trends across models.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
Closed-Source Models											
Generation Models											
|– GPT-4o	**88.24**	97.45	97.17	71.87	64.17	88.11	87.63	93.84	**18.25**	**67.16**	14.59
|– Qwen-Max	84.06	95.62	95.12	76.82	69.82	90.02	88.89	94.45	24.94	64.46	10.60
|– Gemini-2.5-Flash	75.39	91.80	91.20	**85.88**	**75.29**	**91.83**	**89.36**	**95.57**	36.44	58.55	5.01
|– Deepseek-V3	85.67	96.28	96.47	81.57	73.20	89.84	88.48	94.10	21.21	64.27	14.52
Reasoning Models											
|– GPT-o1	72.88	94.48	93.19	77.41	69.53	89.60	88.73	95.11	38.50	56.75	**4.76**
|– Deepseek-R1	86.44	97.55	97.04	68.15	61.81	86.03	85.50	92.47	20.89	57.84	21.27
|– Gemini-2.5-Pro	80.85	**98.30**	**97.94**	75.03	69.06	88.20	87.34	94.01	30.59	61.95	7.46
|– QwQ-32b	74.23	95.14	94.02	78.72	68.08	89.53	88.03	94.82	37.34	54.82	7.84
|– Qwen3-235B-A22B	81.43	96.56	96.02	70.16	63.66	86.31	85.82	93.04	27.70	57.78	14.52
Open-Source Models											
MOE Models											
|– Llama4 Maverick	74.81	92.83	91.52	76.79	66.51	88.47	87.29	93.98	37.53	54.18	8.29
|– Llama4 Scout	75.64	88.26	87.21	67.04	56.34	86.07	86.14	93.37	32.58	52.38	15.04
|– Qwen3-235B-A22B	78.28	94.34	94.34	77.73	68.08	89.70	88.12	94.43	32.33	56.23	11.44
|– Qwen3-30B-A3B	77.83	89.34	88.95	79.50	69.87	89.43	87.79	93.87	29.37	57.58	13.05
Dense Models											
|– Llama3-8B	74.10	83.36	82.01	74.87	64.12	86.72	84.48	92.59	35.03	43.44	21.53
|– Llama3-3B	73.97	83.50	82.07	74.61	63.93	86.76	84.49	92.57	34.51	44.02	21.47
|– Qwen3-8B	74.42	83.45	83.93	82.04	70.07	89.96	87.87	94.00	33.10	55.40	11.50
|– Qwen3-4B	68.38	73.41	74.16	78.30	68.99	88.52	87.46	95.25	40.23	48.59	11.18

Table 6.1: Overall performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity. **Bold boxed entries** highlight the best performance for each metric among closed-source models, while non-bold boxed entries highlight the best performance among open-source models.

Detoxification Accuracy. Most models demonstrate strong performance in removing toxic content. Among generation models, **GPT-4o** achieves the highest S-CLS score (88.24), indicating that its rewrites are most likely to be classified as non-toxic. However, in terms of word-level detoxification, reasoning models such as **Gemini-2.5-Pro** and **Deepseek-R1** outperform generation models, achieving W-Clean scores of 98.30 and 97.55 respectively. This suggests that reasoning models are better at explicitly removing toxic terms. Reasoning models such as **Deepseek-R1** and **QwQ-32B** reveals an interesting trade-off. These models demonstrate strong abilities in identifying toxic triggers and understanding the rewriting intent of preserving emotional polarity. This explains their higher W-Clean and S-Clean scores. However, due to their inclination to generate emotionally intense rewrites—possibly as a result of faithfully preserving tone—these models are more likely to be flagged as still toxic under S-CLS evaluation. This contrast highlights their sensitivity to tone but relative rigidity in emotional modulation.

We observe a nuanced performance gap, comparing open-source and closed-source models. Closed-source commercial models, particularly GPT-4o, Deepseek-V3, and Qwen-Max, consistently achieve higher scores on Detox. Acc., indicating superior control over overall output toxicity. Notably, large open-source MoE models such as **Qwen3-235B-A22B** and **Llama4 Maverick** achieve W-Clean and S-Clean scores

comparable to those of closed models, suggesting that they are similarly effective at eliminating explicit toxic terms. Among open-source models, we find a strong correlation between model scale and detoxification quality: larger models tend to perform better in both sentence-level and word-level detoxification. This trend is also evident within the MoE models. For instance, **Llama4-Maverick** (400B total parameters, 17B active) consistently outperforms **Llama4-Scout** (109B total, 17B active), suggesting larger expert pools provide better representation capacity under the same activation budget.

Fluency and Content Preservation. **Gemini-2.5-Flash** ranks highest in fluency metrics, including BLEU (85.88), ChrF++ (75.29), and BERTScore-F1 (91.83), with **Qwen-Max** and **Deepseek-V3** following closely. These models produce rewrites that are highly natural and grammatically well-formed. Notably, the COMET and content preservation scores largely align with fluency, suggesting that high-quality generation also correlates with better semantic fidelity. Among open-source models, **Qwen3-8B** and **Qwen3-30B-A3B** show strong fluency and preservation, rivaling closed-source systems. Interestingly, we observe that the gap between closed-source and open-source models is minimal in fluency and content preservation. While closed models still lead in detoxification metrics, several open-source models—especially **Qwen3-8B** and **Qwen3-3B**—match or even outperform their commercial counterparts in generation quality. We also find little difference between dense and MoE architectures on these two dimensions. For example, both **Llama4-Maverick** (MoE) and **Llama3-8B** (dense) yield comparable fluency scores, indicating that model architecture has limited impact on fluency and content preservation performance. These results indicate that modern LLMs—even at moderate scales—have largely mastered the ability to produce fluent, semantically faithful rewrites. The true challenge lies not in rewriting per se, but in understanding subtle toxic expressions, interpreting context, and performing sentiment-preserving detoxification.

6.5.5 Sentiment Polarity Consistency Analysis

Maintaining the emotional tone of the original toxic sentence is a crucial goal of our task. In Table 6.1, generation models like **GPT-4o** and **Qwen-Max** strike a good balance between detoxification and emotional preservation, achieving relatively high neutral rates (67.16 and 64.46). In contrast, dense open-source models such as **Llama3-8B** and **Llama3-3B** exhibit higher polite rates (21.53 and 21.47), indicating a tendency to over-sanitize the emotional content. Across the results, we observe that **closed-source commercial models** tend to achieve significantly higher neutral rates compared to open-source models. Most open-source models fall below 58%. Additionally, within the open-source group, **larger MoE models consistently outperform smaller dense models** in polarity consistency. For example, **Qwen3-30B-A3B** (MoE) yields a neutral rate of 57.58% with only 13.05% polite outputs, whereas **Llama3-8B** (dense) produces polite rewrites in 21.53% of cases. These findings suggest that sentiment-preserving detoxification remains a highly challenging task that requires both lexical-level toxicity detection and contextual understanding of emotional intent. Furthermore, models must overcome their tendency to generate overly polite, customer-service-style rewrites, especially in ambiguous or emotionally charged contexts. Besides, reasoning models such as **GPT-o1** and **Gemini-2.5-Pro** demonstrate a deeper understanding of the rewriting goal by producing emotionally expressive, context-sensitive outputs. As a result, they achieve lower polite rates (4.76% and 7.46%, respectively), indicating reduced over-sanitization. However, their tendency to generate emotionally intense rewrites leads

to higher toxicity rates in the sentiment classifier output, consistent with their lower S-CLS scores. These results highlight that effective sentiment-preserving detoxification requires more nuanced modeling of emotional tone, intent, and pragmatic balance. It remains an open challenge, particularly for smaller and open-source models, and calls for future work on targeted emotional style control.

6.5.6 Performance Metrics of Different Scenarios

This section provides the detailed single-sentence evaluation scores in Table 6.2. The remaining scenario-specific tables are placed next to the corresponding analyses below, so the quantitative evidence follows the discussion more closely.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
Closed-Source Models											
Generation Models											
|– GPT-4o	**96.58**	99.52	99.39	41.25	31.46	83.69	86.1	90.73	**6.23**	**79.00**	14.77
|– Qwen-Max	93.04	97.81	97.68	48.14	37.66	86.06	87.68	91.44	12.58	77.29	10.13
|– Gemini-2.5-Flash	87.91	96.57	96.09	**66.97**	**53.91**	**88.37**	**88.28**	**93.08**	24.42	69.35	**6.23**
|– Deepseek-V3	95.60	**99.62**	**99.63**	55.21	43.87	85.25	86.67	90.85	8.30	77.29	14.41
Reasoning Models											
|– GPT-o1	90.48	96.28	95.97	42.88	31.88	84.41	87.12	92.20	19.66	74.11	**6.23**
|– Deepseek-R1	95.85	98.28	98.17	34.19	28.16	80.12	82.60	88.47	9.40	65.32	25.27
|– Gemini-2.5-Pro	94.51	98.57	98.53	43.30	36.84	82.57	84.91	90.52	13.92	75.46	10.62
|– QwQ-32b	85.71	97.04	96.70	52.02	39.16	85.09	86.51	91.97	26.25	65.08	8.67
|– Qwen3-235B-A22B	91.21	98.67	98.41	30.60	24.32	79.74	82.72	88.90	17.70	64.59	17.70
Open-Source Models											
MOE Models											
|– Llama4 Maverick	83.27	91.99	90.84	49.09	37.03	83.98	85.48	90.85	25.52	63.49	10.99
|– Llama4 Scout	86.20	95.52	94.99	43.97	32.21	82.92	84.93	90.44	25.40	66.42	8.18
|– Qwen3-235B-A22B	89.26	82.84	81.20	63.32	46.51	85.28	85.05	91.08	19.54	69.11	11.36
|– Qwen3-30B-A3B	90.72	82.75	81.32	64.26	47.10	85.33	84.99	91.03	15.38	72.04	12.58
Dense Models											
|– Llama3-8B	76.07	97.62	97.56	48.42	35.95	85.76	86.81	91.42	35.41	52.99	11.60
|– Llama3-3B	76.19	96.85	96.21	49.08	34.60	84.98	86.24	90.42	34.07	54.33	11.60
|– Qwen3-8B	89.87	94.85	94.26	41.84	29.90	83.02	85.41	90.38	15.51	71.06	13.43
|– Qwen3-4B	78.27	84.46	82.66	60.47	42.93	86.81	86.91	92.81	30.77	61.29	7.94

Table 6.2: Single-sentence performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity. Bold boxed entries highlight the best performance for each metric among closed-source models, while non-bold boxed entries highlight the best performance among open-source models.

6.5.7 Challenges in Perturbation Toxic Rewrite

While models perform well on standard single-sentence rewrite, we observe significant degradation in both **emoji-induced** and **homophone-based** settings (shown in Figure 6.10), revealing key limitations in handling *implicit and structurally masked toxicity*. Detoxification accuracy declines sharply in these subsets. Models that perform similarly on standard inputs diverge substantially when facing emoji and homophone perturbations, suggesting divergent capacities in interpreting obfuscated toxicity. In homophones, content preservation drops slightly due to necessary substitutions that alter surface form. Sentiment polarity also deteriorates. Neutral output rates fall, while toxicity increases—without a corresponding rise in politeness—suggesting that

models fail to resolve deeper aggression embedded in sarcasm (emojis) or veiled insults (homophones). These findings expose a bottleneck in LLMs' ability to handle *covert toxicity*, where emotion, intent, and context interact beneath surface-level fluency. Detailed results on emoji-induced and homophone-based toxicity rewriting are provided in Tables 6.3– 6.4.

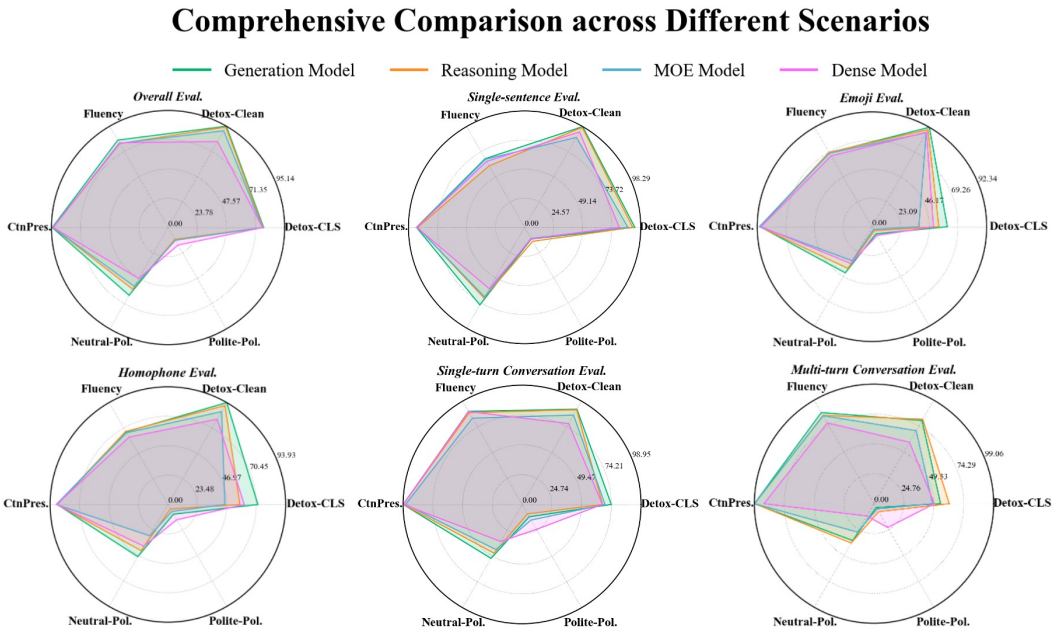


Figure 6.10: Comparison of four model variants (Generation, Reasoning, MOE, and Dense) across different evaluation scenarios: overall, single-sentence, emoji, homophone, single-turn conversation, and multi-turn conversation. Each chart visualizes performance on six metrics: Detox-CLS, Detox-Clean, Fluency, Content Preservation, Neutral Polarity, and Polite Polarity.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite↓
Closed-Source Models											
Generation Models											
|– GPT-4o	75.51	95.31	93.88	47.55	39.09	82.71	82.75	89.20	40.82	51.02	8.16
|– Qwen-Max	59.18	92.19	89.80	58.01	52.23	85.05	84.51	90.10	53.06	38.78	8.16
|– Gemini-2.5-Flash	32.65	85.94	81.63	**69.79**	52.35	**87.21**	**85.94**	91.62	73.47	26.53	**0.00**
|– Deepseek-V3	75.51	**100.00**	**100.00**	56.12	44.73	85.34	85.69	88.97	34.69	55.10	10.20
Reasoning Models											
|– GPT-o1	36.73	87.50	83.67	68.03	**54.39**	87.03	85.36	**92.31**	69.39	28.57	2.04
|– Deepseek-R1	**77.55**	93.75	91.84	44.67	33.58	81.94	81.30	86.60	**32.65**	**59.18**	8.16
|– Gemini-2.5-Pro	**77.55**	96.88	95.92	58.07	50.16	84.97	85.15	89.34	46.94	48.98	4.08
|– QwQ-32b	24.49	85.94	81.63	66.63	52.70	85.90	83.70	91.97	81.63	16.33	2.04
|– Qwen3-235B-A22B	53.06	95.31	93.88	62.04	46.20	86.02	83.73	91.56	57.14	40.82	2.04
Open-Source Models											
MOE Models											
|– Llama4 Maverick	34.69	85.94	81.63	58.24	39.35	83.85	83.56	89.84	69.39	30.61	0.00
|– Llama4 Scout	36.73	90.62	87.76	55.99	38.28	85.00	83.94	90.91	67.35	30.61	2.04
|– Qwen3-235B-A22B	34.69	90.62	87.76	64.38	50.82	86.51	84.03	91.36	71.43	26.53	2.04
|– Qwen3-30B-A3B	44.90	89.06	85.71	64.19	47.37	85.32	82.51	90.27	55.10	38.78	6.12
Dense Models											
|– Llama3-8B	40.82	89.06	85.71	58.75	47.12	82.87	79.59	88.38	67.35	26.53	6.12
|– Llama3-3B	40.82	89.06	85.71	58.09	46.27	82.69	79.89	88.23	67.35	26.53	6.12
|– Qwen3-8B	57.14	90.62	87.76	45.32	34.57	81.51	82.73	89.22	44.90	46.94	8.16
|– Qwen3-4B	59.18	89.06	85.71	59.33	46.41	84.25	82.82	91.01	51.02	36.73	12.24

Table 6.3: Emoji performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity. Bold boxed entries highlight the best performance for each metric among closed-source models, while non-bold boxed entries highlight the best performance among open-source models.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite↓
Closed-Source Models											
Generation Models											
|– GPT-4o	**82.05**	**98.15**	**97.44**	45.42	36.28	81.76	82.91	86.14	41.03	41.03	17.95
|– Qwen-Max	69.23	94.44	92.31	52.62	47.48	83.89	83.10	88.76	**30.77**	**61.54**	7.69
|– Gemini-2.5-Flash	56.41	88.89	84.62	**66.07**	**49.40**	**85.65**	84.22	**90.13**	58.97	41.03	**0.00**
|– Deepseek-V3	79.49	**98.15**	**97.44**	57.70	44.93	84.40	**84.88**	88.14	41.03	48.72	10.26
Reasoning Models											
|– GPT-o1	46.15	88.89	84.62	61.75	48.20	85.26	84.29	89.65	66.67	30.77	2.56
|– Deepseek-R1	74.36	94.44	92.31	52.06	37.80	82.99	81.53	86.73	38.46	53.85	7.69
|– Gemini-2.5-Pro	66.67	96.30	94.87	59.21	47.17	84.53	84.40	89.25	43.59	53.85	2.56
|– QwQ-32b	41.03	90.74	87.18	56.96	45.66	83.72	80.72	88.68	71.79	25.64	2.56
|– Qwen3-235B-A22B	58.97	94.44	92.31	60.87	44.50	84.09	81.88	88.56	46.15	48.72	5.13
Open-Source Models											
MOE Models											
|– Llama4 Maverick	46.15	85.19	79.49	56.27	37.80	82.30	81.38	88.40	61.54	33.33	5.13
|– Llama4 Scout	35.90	88.89	84.62	52.43	39.82	83.52	81.59	89.34	76.92	23.08	0.00
|– Qwen3-235B-A22B	48.72	92.59	89.74	56.11	41.60	83.78	81.84	88.82	64.10	25.64	10.26
|– Qwen3-30B-A3B	51.28	85.19	79.49	63.80	47.49	85.20	82.79	90.14	56.41	33.33	10.26
Dense Models											
|– Llama3-8B	58.97	81.48	76.92	48.82	33.90	78.63	74.07	86.18	46.15	38.46	15.38
|– Llama3-3B	56.41	81.48	76.92	51.25	40.68	80.76	76.74	87.50	48.72	41.03	10.26
|– Qwen3-8B	69.23	85.19	82.05	45.72	33.92	81.62	81.20	89.78	41.03	41.03	17.95
|– Qwen3-4B	58.97	75.93	69.23	56.23	43.37	83.59	82.26	91.69	53.85	33.33	12.82

Table 6.4: Homophone performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity. Bold boxed entries highlight the best performance for each metric among closed-source models, while non-bold boxed entries highlight the best performance among open-source models.

6.5.8 Challenges in Conversation Toxic Rewrite

Detoxifying toxic language in conversations poses unique challenges due to context dependence and emotional continuity. Comparing single-turn and multi-turn detoxification, we find that performance drops sharply in multi-turn settings. Detoxification accuracy declines across the board in multi-turn dialogue. Top models like GPT-4o and Qwen-Max see S-CLS scores fall below 56%, and both sentence- and word-level detox metrics degrade—indicating difficulties in tracing and neutralizing toxicity that unfolds across turns. This highlights a key limitation in current LLMs' ability to align detoxification with dialogue structure and intent. Fluency and content preservation remain stable, with most models generating coherent outputs. However, smaller dense models show minor drops in fluency, suggesting limited capacity to manage long-range discourse. Sentiment polarity control weakens in multi-turn scenarios. Toxicity rates rise significantly, while neutral output rates fall—without a rise in polite rewrites—revealing that models fail to neutralize cumulative or reactive toxicity rather than merely oversanitizing. Overall, multi-turn dialogue is the most difficult setting, where toxicity often accumulates contextually or emotionally. These findings suggest that successful detoxification in dialogue requires discourse-level reasoning and pragmatic awareness beyond sentence rewriting. Detailed results are provided in Tables 6.5– 6.6.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
Closed-Source Models											
Generation Models											
|– GPT-4o	**80.49**	94.45	96.00	84.54	75.23	94.56	90.38	98.56	**28.78**	**55.93**	15.28
|– Qwen-Max	76.59	91.85	94.16	88.47	80.01	95.76	91.21	98.90	36.59	51.38	12.03
|– Gemini-2.5-Flash	64.88	83.19	87.19	**93.12**	82.77	**96.98**	**91.45**	**99.37**	46.50	49.27	4.23
|– Deepseek-V3	74.80	90.12	93.14	91.86	**82.79**	96.31	91.34	98.95	34.15	50.08	15.77
Reasoning Models											
|– GPT-o1	55.12	92.37	93.48	90.78	81.40	96.67	91.44	99.32	56.91	39.67	**3.41**
|– Deepseek-R1	75.93	96.88	97.60	82.45	72.84	94.09	89.92	98.31	33.66	48.13	18.21
|– Gemini-2.5-Pro	65.20	**98.09**	**98.17**	89.18	80.31	95.93	90.92	99.08	48.78	47.15	4.07
|– QwQ-32b	65.53	91.33	93.94	88.49	77.49	95.88	90.89	99.01	45.69	46.99	7.32
|– Qwen3-235B-A22B	73.17	93.07	94.74	86.26	76.80	94.87	90.31	98.62	36.42	51.38	12.20
Open-Source Models											
MOE Models											
|– Llama4 Maverick	72.52	85.58	84.88	88.13	77.09	94.99	90.37	98.59	35.28	41.95	22.76
|– Llama4 Scout	66.50	90.73	88.78	75.18	63.93	90.13	88.16	97.42	46.83	43.58	9.59
|– Qwen3-235B-A22B	69.92	83.75	83.74	80.53	70.86	89.89	85.57	95.81	42.60	44.88	12.52
|– Qwen3-30B-A3B	66.99	83.98	83.58	80.70	70.92	90.02	85.73	95.94	42.11	42.76	15.12
Dense Models											
|– Llama3-8B	75.77	91.99	92.20	89.24	78.67	95.33	90.66	98.78	30.24	34.31	35.45
|– Llama3-3B	75.28	82.61	82.11	90.17	80.66	95.59	90.57	98.70	30.89	33.98	35.12
|– Qwen3-8B	57.40	71.17	72.52	91.71	81.06	96.39	90.92	99.15	52.52	38.21	9.27
|– Qwen3-4B	58.54	62.01	65.20	89.05	78.03	94.66	89.90	98.82	48.78	35.45	15.77

Table 6.5: Single-turn conversation performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity. Bold boxed entries highlight the best performance for each metric among closed-source models, while non-bold boxed entries highlight the best performance among open-source models.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
Closed-Source Models											
Generation Models											
|– GPT-4o	52.94	87.50	73.53	77.90	70.65	92.94	87.30	98.90	58.82	38.24	2.94
|– Qwen-Max	55.88	86.25	70.59	88.61	80.67	95.94	**88.86**	99.10	64.71	32.35	2.94
|– Gemini-2.5-Flash	47.06	86.25	76.47	90.16	79.82	95.78	88.23	99.06	64.71	32.35	2.94
|– Deepseek-V3	64.71	82.50	73.53	**90.38**	**83.46**	**96.10**	88.55	99.16	55.88	38.24	5.88
Reasoning Models											
|– GPT-o1	52.94	91.25	79.41	87.32	81.14	95.41	88.55	**99.18**	82.35	17.65	**0.00**
|– Deepseek-R1	**76.47**	92.50	85.29	78.60	73.32	91.85	85.93	98.21	**29.41**	**55.88**	14.71
|– Gemini-2.5-Pro	55.88	**98.75**	**97.06**	80.67	74.39	93.02	87.49	98.74	64.71	32.35	2.94
|– QwQ-32b	64.71	93.75	85.29	84.93	76.23	93.60	87.37	98.85	50.00	38.24	11.76
|– Qwen3-235B-A22B	61.76	91.25	85.29	79.62	71.16	92.79	86.96	98.72	47.06	44.12	8.82
Open-Source Models											
MOE Models											
|– Llama4 Maverick	41.18	72.50	58.82	77.85	67.26	92.55	87.45	98.55	67.65	26.47	5.88
|– Llama4 Scout	50.00	85.00	70.59	76.89	67.48	93.03	86.87	99.11	73.53	20.59	5.88
|– Qwen3-235B-A22B	61.76	81.25	70.59	82.37	75.16	94.27	87.08	99.01	61.76	29.41	8.82
|– Qwen3-30B-A3B	41.18	67.50	52.94	90.56	83.14	96.02	88.18	99.32	67.65	32.35	0.00
Dense Models											
|– Llama3-8B	61.76	82.50	70.59	58.59	60.22	79.14	70.28	84.29	52.94	8.82	38.24
|– Llama3-3B	64.71	85.00	73.53	48.87	54.63	74.95	65.69	80.75	47.06	5.88	47.06
|– Qwen3-8B	41.18	61.25	38.24	92.23	86.01	96.67	88.52	99.64	79.41	17.65	2.94
|– Qwen3-4B	32.35	38.75	20.59	91.64	84.36	96.32	88.06	99.60	82.35	14.71	2.94

Table 6.6: Multi-turn conversation performance metrics of various models across detoxification, fluency, content preservation, and sentiment polarity. Bold boxed entries highlight the best performance for each metric among closed-source models, while non-bold boxed entries highlight the best performance among open-source models.

We further examine performance from simple to complex scenarios, and quantify how each added complexity affects model performance across metrics. As shown in Table 6.7, both Detox. Acc. and Sentiment Polarity scores decline with increasing task difficulty. Interestingly, fluency improves in dialogue scenarios, possibly due to richer contextual flow, and content preservation also benefits.

Setting	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
Qwen-Max											
single-sentence	93.04	97.81	97.68	48.14	37.66	86.06	87.68	91.44	12.58	77.29	10.13
+ single-turn conv.	85.98	96.15	95.96	76.52	69.41	90.22	89.19	94.64	22.87	66.18	10.95
	(-7.06)	(-1.66)	(-1.72)	(+28.38)	(+31.75)	(+4.16)	(+1.51)	(+3.20)	(+10.29)	(-11.11)	(+0.82)
+ single-turn conv. + homo	85.54	96.11	95.86	76.29	69.18	90.05	89.03	94.48	23.08	66.06	10.86
	(-0.44)	(-0.04)	(-0.10)	(-0.23)	(-0.23)	(-0.17)	(-0.16)	(-0.16)	(+0.21)	(-0.12)	(-0.09)
+ single-turn conv. + homo	84.69	95.98	95.66	76.07	68.95	89.89	88.89	94.34	24.05	65.18	10.78
+ emoji	(-0.85)	(-0.13)	(-0.20)	(-0.22)	(-0.23)	(-0.16)	(-0.14)	(-0.14)	(+0.97)	(-0.88)	(-0.08)
+ single-turn conv. + homo	84.06	95.62	95.12	76.82	69.82	90.02	88.89	94.45	24.94	64.46	10.60
+ emoji + multi-turn conv.	(-0.63)	(-0.36)	(-0.54)	(+0.75)	(+0.87)	(+0.13)	(±0.00)	(+0.11)	(+0.89)	(-0.72)	(-0.18)
GPT-o1											
single-sentence	90.48	96.28	95.97	42.88	31.88	84.41	87.12	92.20	19.66	74.11	6.23
+ single-turn conv.	75.31	95.01	94.07	76.97	69.01	89.67	88.97	95.25	35.63	59.34	5.02
	(-5.17)	(-1.27)	(-1.90)	(+34.09)	(+37.13)	(+5.26)	(+1.85)	(+3.05)	(+15.97)	(-14.77)	(-1.21)
+ single-turn conv. + homo	74.54	94.84	93.82	76.85	68.80	89.55	88.85	95.11	36.46	58.59	4.96
	(-0.77)	(-0.17)	(-0.25)	(-0.12)	(-0.21)	(-0.12)	(-0.12)	(-0.14)	(+0.83)	(-0.75)	(-0.06)
+ single-turn conv. + homo	73.32	94.61	93.50	76.76	68.59	89.47	88.73	95.02	37.52	57.62	4.86
+ emoji	(-1.22)	(-0.23)	(-0.32)	(-0.09)	(-0.21)	(-0.08)	(-0.12)	(-0.09)	(+1.06)	(-0.97)	(-0.10)
+ single-turn conv. + homo	72.88	94.48	93.19	77.41	69.53	89.60	88.73	95.11	38.50	56.75	4.76
+ emoji + multi-turn conv.	(-0.44)	(-0.13)	(-0.31)	(+0.65)	(+0.94)	(+0.13)	(±0.00)	(+0.09)	(+0.98)	(-0.87)	(-0.10)
Qwen3-235B-A22B											
single-sentence	89.26	97.62	97.56	48.42	35.95	85.76	86.81	91.42	19.54	69.11	11.36
+ single-turn conv.	80.96	95.06	95.26	77.67	67.92	89.86	88.46	94.57	29.43	58.72	11.85
	(-8.30)	(-2.56)	(-2.30)	(+29.25)	(+31.97)	(+4.10)	(+1.65)	(+3.15)	(+9.89)	(-10.39)	(+0.49)
+ single-turn conv. + homo	80.11	94.99	95.11	77.54	67.70	89.70	88.28	94.42	30.35	57.84	11.81
	(-0.85)	(-0.07)	(-0.15)	(-0.13)	(-0.22)	(-0.16)	(-0.18)	(-0.15)	(+0.92)	(-0.88)	(-0.04)
+ single-turn conv. + homo	78.65	94.86	94.88	77.43	67.47	89.60	88.15	94.32	31.67	56.83	11.50
+ emoji	(-1.46)	(-0.13)	(-0.23)	(-0.11)	(-0.23)	(-0.10)	(-0.13)	(-0.10)	(+1.32)	(-1.01)	(-0.31)
+ single-turn conv. + homo	78.28	94.34	94.34	77.73	68.08	89.70	88.12	94.43	32.33	56.23	11.44
+ emoji + multi-turn conv.	(-0.37)	(-0.52)	(-0.54)	(+0.30)	(+0.61)	(+0.10)	(-0.03)	(+0.11)	(+0.66)	(-0.60)	(-0.06)
Qwen3-8B											
single-sentence	89.87	94.85	94.26	41.84	29.90	83.02	85.41	90.38	15.51	71.06	13.43
+ single-turn conv.	75.94	84.09	84.94	77.96	68.24	88.75	87.77	94.14	31.38	56.97	11.65
	(-13.93)	(-10.76)	(-9.32)	(+36.12)	(+38.34)	(+5.73)	(+2.36)	(+3.76)	(+15.87)	(-14.09)	(-1.78)
+ single-turn conv. + homo	75.76	84.12	84.86	77.70	67.93	88.56	87.60	94.03	31.64	56.55	11.81
	(-0.18)	(+0.03)	(-0.08)	(-0.26)	(-0.31)	(-0.19)	(-0.17)	(-0.11)	(+0.26)	(-0.42)	(+0.16)
+ single-turn conv. + homo	75.16	84.32	84.95	77.35	67.55	88.34	87.44	93.87	32.06	56.24	11.70
+ emoji	(-0.60)	(+0.20)	(+0.09)	(-0.35)	(-0.38)	(-0.22)	(-0.16)	(-0.16)	(+0.42)	(-0.31)	(-0.11)
+ single-turn conv. + homo	74.42	83.45	83.93	82.04	70.07	89.96	87.87	94.00	33.10	55.40	11.50
+ emoji + multi-turn conv.	(-0.74)	(-0.87)	(-1.02)	(+4.69)	(+2.52)	(+1.62)	(+0.43)	(+0.13)	(+1.04)	(-0.84)	(-0.20)

Table 6.7: Progressive breakdown of model performance with incrementally added task complexity.

6.5.9 Human Preference Analysis

To assess whether sentiment-preserving rewrites are more aligned with human expectations, we conducted a preference study over 100 sampled examples spanning all five categories: sentence-level (52), emoji (3), homophone (3), single-turn (40), and multi-turn (2). For each example, annotators were shown two rewrites of the same toxic input: one neutral (preserving sentiment) and one polite (over-sanitized), generated by two competitive models—GPT-4o and Deepseek-V3—randomized in order. Annotators were asked to select the version that better preserved the original user intent while removing toxicity.

As shown in Figure 6.11, out of 100 comparisons, the sentiment-preserving (neutral) rewrite was preferred in 79 cases, while the polite rewrite was chosen in 21 cases. This

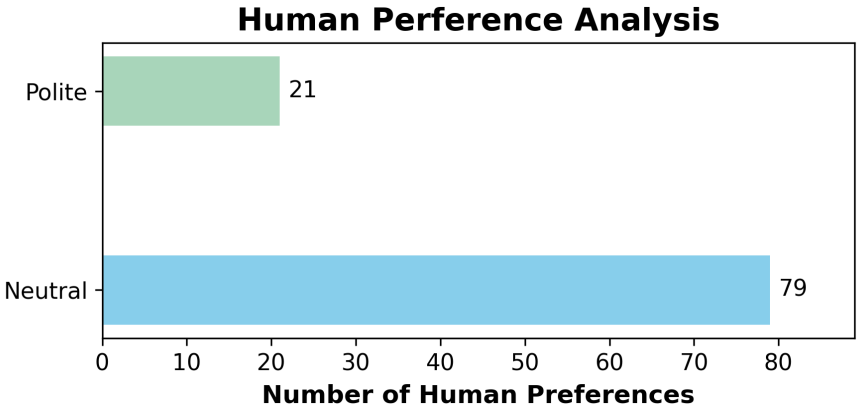


Figure 6.11: Human preference comparison between sentiment-preserving (neutral) and over-polite detoxification rewrites. Annotators significantly preferred neutral rewrites (79%) over polite ones (21%).

indicates a strong human preference for rewrites that retain emotional tone rather than overly formal rephrasings, reinforcing the core motivation of ToxiRewriteCN.

To better assess model performance and the reliability of automatic metrics, we include a **human rewrite baseline**. Two native Chinese speakers—who were neither annotators nor familiar with the task—were asked to rewrite 30 sampled toxic sentences covering all five categories, using the same prompt format as used for LLMs. Their rewrites were scored using the same metrics. As shown in Table 6.8, human rewrites significantly outperform LLMs across detoxification accuracy (S-CLS), content preservation, and especially sentiment control—achieving high neutral rates with minimal residual toxicity.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
GPT-4o	83.33	100.00	100.00	73.97	66.61	86.44	85.77	91.87	36.67	53.33	10.00
Deepseek-R1	90.00	100.00	100.00	77.56	71.32	86.73	84.87	90.61	36.67	46.67	16.67
Annonator1	96.67	100.00	100.00	83.75	73.24	90.99	87.98	92.61	3.33	90.00	6.67
Annonator2	93.34	100.00	100.00	85.64	77.65	91.03	87.73	93.35	3.33	83.33	13.33

Table 6.8: Performance between LLMs and human annotators on detoxification, fluency, content preservation, and sentiment polarity.

6.5.10 Fine-tuning with 1K Samples

To complement our main evaluation across a broad range of models—including commercial and open-source systems, generation- and reasoning-oriented models, as well as both dense and MoE architectures—we further trained a task-specific baseline to better assess the potential of fine-grained supervision on our benchmark.

We constructed a fine-tuning setup using only our proposed ToxiRewriteCN. A sample of 1,000 toxic–rewrite pairs (preserving original category distribution) was used for training, with the remaining 556 examples reserved for evaluation.

We fine-tuned LLaMA3-8B using the reasoning traces generated by Deepseek-R1 as supervision signals. Table 6.9 demonstrates that the resulting model, LLaMA3-8B_1K, achieved substantial improvements across multiple metrics. Compared to its original version (LLaMA3-8B), S-CLS improved by **10.79 points** (from 73.02 to 83.81), and the proportion of neutral sentiment rewrites increased by **21.76 points**, outperforming even Deepseek-R1 (**66.73% vs. 58.63%**).

These results demonstrate that even with limited, high-quality supervision (1K samples), our benchmark enables meaningful gains in both detoxification accuracy and sentiment control—suggesting its utility for training robust detoxification models under low-resource conditions.

Model	Detox. Acc.			Fluency				CntPres.↑	Sentiment Polarity		
	S-CLS↑	W-Clean↑	S-Clean↑	BLEU↑	ChrF++↑	BS_F1↑	COM.↑		Toxic↓	Neutral↑	Polite ↓
Deepseek-R1	**87.05**	99.72	99.64	68.49	61.74	86.24	85.79	92.57	**21.22**	58.63	20.14
Llama3-8B	73.02	99.02	98.74	75.79	63.30	86.87	84.66	**93.04**	36.33	42.09	21.58
Llama3-8B_1K	83.81	**100.00**	**100.00**	**76.83**	**66.50**	**87.39**	**85.81**	92.77	21.94	**63.85**	**14.21**

Table 6.9: Performance of fine-tuning LLaMA3-8B with 1K samples. LLaMA3-8B_1K yields clear gains over the base model, notably +10.79 on S-CLS and +21.76 on neutral polarity, surpassing Deepseek-R1.

6.6 Conclusion

We introduce ToxiRewriteCN, the first Chinese detoxification dataset that explicitly preserves sentiment polarity—an essential yet underexplored aspect in controllable toxic language rewriting. Through a comprehensive evaluation, spanning commercial and open-source models with diverse architectures and parameter scales, we uncover the trade-offs and limitations of current systems in balancing safety and expressive fidelity. Our scenario-level analysis further highlights the unique challenges posed by implicit toxicity from emojis and homophones, as well as contextually emergent toxicity in multi-turn dialogues. We hope this work provides a foundation for future research on sentiment-aware, context-sensitive detoxification in Chinese and other low-resource, high-complexity settings.

My experience is what I agree to attend to.

— William James (1890)

7

Cognition-Inspired Methods for Efficiently Semantic Steering

Contents

Publication Note. This chapter is based on and extends two first-authored publications by the dissertation author: Wang et al., "Probing Large Language Models from A Human Behavioral Perspective," published in NeusymBridge at LREC-COLING 2024; and Wang et al., "CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models," published in Findings of the Association for Computational Linguistics (ACL 2025). For integration into this dissertation, the material has been consolidated and lightly revised for consistency in terminology, formatting, and cross-references. The core contributions and findings remain unchanged.

7.1 Abstract

Large Language Models (LLMs) achieve remarkable performance through pretraining on extensive data. This enables efficient adaptation to diverse downstream tasks. However, the lack of interpretability in their underlying mechanisms limits the ability to effectively steer LLMs for specific applications. In this work, we investigate the intrinsic mechanisms of LLMs from a cognitive perspective using eye movement measures. Specifically, we analyze the layer-wise correlation between human cognitive indicators and LLM representations. Building on these insights, we propose a heuristic approach for selecting the optimal steering layer to modulate LLM semantics. To this end, we introduce an efficient selective layer intervention based on prominent *parameter-efficient fine-tuning* methods, which conventionally adjust either all layers or only the final layer. Additionally, we present an *implicit layer contrastive intervention* during inference to steer LLMs away from toxic outputs. Extensive experiments on natural language understanding, reasoning, and generation tasks, conducted on GPT-2, Llama2-7B, and Mistral-7B, demonstrate the effectiveness and efficiency of our approach. As a model-agnostic framework, it enhances the interpretability of LLMs while improving efficiency for safe deployment.

7.2 Introduction

Large Language Models (LLMs) (Dubey et al., 2024; A Yang et al., 2024; Guo et al., 2025) have demonstrated strong capabilities in natural language understanding and reasoning (Y Zhao et al., 2024; WX Zhao et al., 2025; Wei, Tay, et al., 2022; H Yin et al., 2025; Zeng et al., 2025) through pretraining on large datasets, followed by instruction tuning and alignment with human values (Wei, Bosma, et al., 2022; Ouyang et al., 2022). Consequently, LLMs achieve excellent performance on downstream tasks with fine-tuning. However, their lack of interpretability and transparency limits the development of efficient fine-tuning and inference methods.

To understand intrinsic mechanisms of LLMs, previous work has introduced various interpretability methods, including training linear classifiers as probes on top of hidden representations (Belinkov, 2022), projecting representations into vocabularies (Geva et al., 2022), and intervening in the computation path, such as knowledge neurons (D Dai et al., 2022) and circuits (Conmy et al., 2023; Ghandeharioun et al., 2024). However,

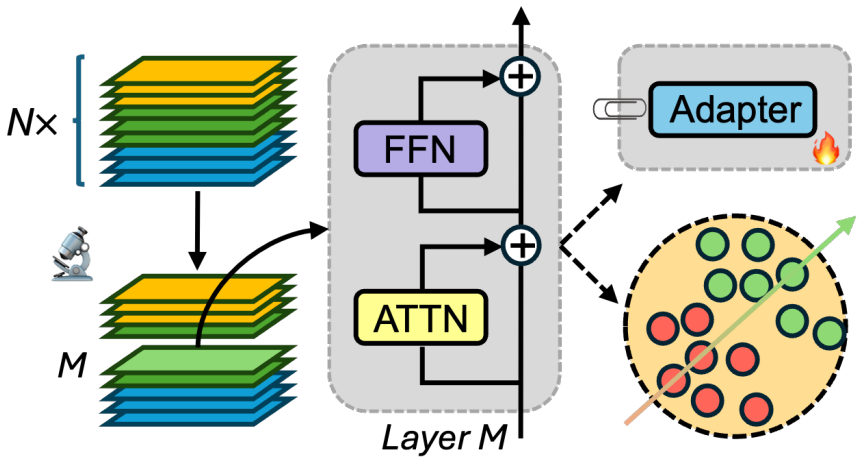


Figure 7.1: Demonstration of CogSteer Intervention. For an N-layer LLM, we first heuristically find the optimal layer M for semantic intervention. The upper block represents an adapter that is fine-tuned and inserted into the frozen layer M. The bottom block illustrates the operation of the attention module in M to steer the semantic direction towards safer outputs during inference.

these methods, which focus on a limited set of predefined classes, concepts, or prompts, have practical limitations in terms of scalability and generalization.

In this work, we introduce a novel interpretability analysis method that leverages eye movement data (Luke and Christianson, 2018; Hollenstein et al., 2020; Colman et al., 2022) collected by cognition researchers to study human reading behaviors. Through correlation experiments, we find that LLM hidden states exhibit a strong correlation with human gaze, peaking in the middle layers. Using eye movement measures such as *fixation* and *regression* (Rayner, 1998) as human-interpretable indicators, we observe a hierarchical progression in LLMs, from initial syntactic and semantic processing to deeper integration and final prediction. Additionally, a comparison of correlation results between natural reading and task-specific reading suggests that the upper layers of LLMs are more capable of reasoning. Furthermore, advanced LLMs such as Llama, which incorporate instruction tuning and reinforcement learning from human feedback (RLHF), demonstrate enhanced reasoning capabilities, even in the middle layers.

LLM intervention (Poth et al., 2023; Dong et al., 2024) refers to fine-tuning or applying inference methods to steer their semantics to align with specific tasks and data distributions. In addition to enhancing the understanding of LLM behavior, our interpretability analysis reveals that different layers serve distinct functions, with middle layers playing a crucial role in deeper syntactic and semantic processing. This enables us to first identify the most suitable layer for intervention, thereby improving task-specific performance. Based on these insights, we propose a heuristic approach for selecting the optimal steering layer for semantic intervention.

To achieve this, we refine prominent Parameter-Efficient Fine-Tuning (PEFT) methods, which traditionally adjust either all layers or only the last layer. As shown in Figure 7.1, our proposed *CogSteer* framework first identifies the most suitable layer M for semantic intervention based on the task. Instead of fine-tuning all layers or only the last layer, our method fine-tunes only the selected layer, enabling LLMs to better adapt to specific tasks and datasets. The number of learnable parameters in *CogSteer* is significantly reduced, requiring only $1/N$ of the parameters in LLMs, thereby

improving parameter efficiency. Furthermore, we propose an *implicit layer contrastive intervention* method during inference, which efficiently identifies and steers semantics toward safer generation directions to evaluate the effectiveness of our proposed selective layer intervention.

Through extensive evaluations across diverse tasks and datasets, we demonstrate that the proposed selective layer intervention method achieves comparable or even superior performance with fewer parameters compared to the full-layer intervention baseline. Specifically, we observed an average absolute improvement of **+1.7** on the GLUE benchmark for Llama2-7B, and an average absolute improvement of **+5.8** on the GLUE benchmark for Mistral-7B, with only **3.1%** of the parameters involved in full-layer fine-tuning. Moreover, in experiments on generation tasks, language toxification (Dementieva et al., 2025), and detoxification (Leong et al., 2023), our method achieves a **+1.85%** improvement in toxification compared to full-layer intervention and a **+13.45%** improvement in detoxification as compared to last-layer intervention.

Our **main contributions** are as follows:

(1) We are the first to propose leveraging eye movement measures to analyze the layer-wise behavior of LLMs. We publicly release the probing code [1] to facilitate further research on interpretability from a cognitive perspective. (2) Through correlation analysis, we demonstrate a hierarchical progression in LLMs and introduce a heuristic steering layer selection method for efficient layer intervention. (3) Extensive experiments validate the effectiveness of our proposed method across various language understanding, reasoning, and generation tasks, contributing to the development of efficient and explainable foundation models.

7.3 Related Work

Interpretability research is essential for uncovering the mechanisms of LLMs and ensuring their safe and trustworthy deployment as foundation models. Various studies have analyzed how knowledge is stored in LLMs (Goldowsky-Dill et al., 2023; Stolfo et al., 2023; Rai et al., 2025; Bills et al., 2023; Geva et al., 2022), focusing on concepts such as knowledge neurons (D Dai et al., 2022) and circuits (Conmy et al., 2023; Y Yao et al., 2025). Moreover, **Wang**, Li, et al. (2024) reveal that GPT-2 predicts tokens more similarly to humans than shallow language models but lacks engagement with LLMs. Oh and Schuler (2023) analyzes how model scale and training data influence surprisal-based predictions of reading time. Gao et al. (2023) investigates the alignment between model and human attention. Unlike previous works, our method investigates the interpretability of LLMs through human-interpretable indicators based on eye movement theory, enabling a more precise understanding and control of model behavior (see § 7.4 and 7.5.1).

Parameter-Efficient Semantic Steering PEFT (Han et al., 2024; **Wang et al.**, 2023), including Adapter (Houlsby et al., 2019; Poth et al., 2023) and LoRA (Hu et al., 2022), has gained popularity due to its ability to maintain a large number of frozen parameters in LLMs for generality while introducing only a small number of trainable parameters per task. Our method for semantic steering via fine-tuning enhances PEFT methods by incorporating a selective layer intervention strategy, reducing computational costs

1. https://github.com/Ethanscuter/CogSteer

while achieving superior performance (see § 7.5.2). Furthermore, steering semantics during inference offers an even more efficient approach. Contrastive decoding (XL Li et al., 2023; Sennrich et al., 2024; **Wang**, Pan, Ding, and Biemann, 2024) guides the generation process by comparing two output distributions. In contrast, our proposed implicit layer contrastive intervention efficiently identifies and steers the semantics of LLMs toward safe directions during inference (see § 7.5.3).

7.4 From Human Gaze to LLM Behavior

Recent studies (Geva et al., 2021; Schuster et al., 2022) suggest that feed-forward networks (FFNs) function similarly to neural memory networks, capturing syntactic and semantic features as well as factual knowledge (Chuang et al., 2024; N Zhang et al., 2024). Meanwhile, research in cognitive science (Rayner, 1998) has shown that eye movement measures provide insights into the time required for human readers to process syntax, semantics, and integrate information. Motivated by these findings, we leverage eye movement measures to analyze their correlation with the hidden states of FFNs across different layers of LLMs.

Models. Correlation studies are conducted on the GPT-2 model (Radford et al., 2019) at different sizes (12-layer small, 24-layer medium, 36-layer large) and the 32-layer Llama2-7B (Touvron, Lavril, et al., 2023). Both GPT-2 and Llama2-7B use a decoder-only transformer architecture. Comparing earlier LLMs, such as GPT-2, with more advanced models like Llama2 provides valuable insights into their similarities and differences while also enhancing the robustness and generalizability of our findings.

Eye-movement Experiments and Datasets. We conduct a correlation analysis under two experimental conditions: natural reading and task-specific reading. For natural reading, we use the Provo (Luke and Christianson, 2018), GECO (Colman et al., 2022), and ZuCo 2.0 (Hollenstein et al., 2020) datasets. The ZuCo 2.0 dataset also includes task-specific reading experiments. In task-specific reading, participants are required to determine whether a specific relation type is present in a sentence. Relation detection is a high-level semantic and reasoning task that involves complex cognitive processing.

Correlation Analysis. Let S_j denote the j-th sentence, consisting of n_j words $w_1, w_2, \ldots, w_{n_j}$. For each word w_i in sentence S_j, we consider five eye movement measures: $e_i^{(k)}$, $k \in \{sfd, ffd, gd, trt, gpt\}$, where each measure represents a scalar value. The hidden state at layer l of the LLM for word w_i is denoted as $\mathbf{h}_{l,i} \in \mathbb{R}^d$, where d is the dimensionality of the hidden states.

To analyze the relationship between eye movement measures and LLM hidden states, we compute the *Pearson correlation* between each eye movement measure and the corresponding hidden states at each layer. Specifically, we concatenate hidden states and eye movement measures across all words in the dataset and apply *principal component analysis* to obtain a scalar representation of hidden states, ensuring alignment with eye movement measures. The correlation is then defined as:

$$\rho_{l,k} = \frac{\sum_{i=1}^{n_{total}} \left(h_{l,i} - \bar{h}_l\right)\left(e_i^{(k)} - \bar{e}^{(k)}\right)}{\sqrt{\sum_{i=1}^{n_{total}} \left(h_{l,i} - \bar{h}_l\right)^2} \sqrt{\sum_{i=1}^{n_{total}} \left(e_i^{(k)} - \bar{e}^{(k)}\right)^2}}, \tag{7.1}$$

where $h_{l,i}$ represents the processed hidden states aligned with the eye movement measures, and $\bar{h}_l$ and $\bar{e}^{(k)}$ are their respective means.

Results and Finding 1. Based on the correlation calculation in Eq. 7.1, we first analyze the correlation results for the natural reading task. To facilitate interpretation, we divide the layers of both models into three equal groups: *premature*, *middle*, and *mature*. The tripartite division of layers into premature, middle, and mature buckets is motivated by recent interpretability studies (Geva et al., 2021; Schuster et al., 2022; Chuang et al., 2024; N Zhang et al., 2024), which consistently observe a coarse-grained functional structure in LLMs: lower layers predominantly capture syntactic patterns, middle layers encode semantic and contextual information, and upper layers are often involved in reasoning and factual retrieval. Figure 7.2 (*upper panels*) presents the correlation results between various eye movement measures and the hidden states of the LLMs, illustrating how these values evolve across layers.

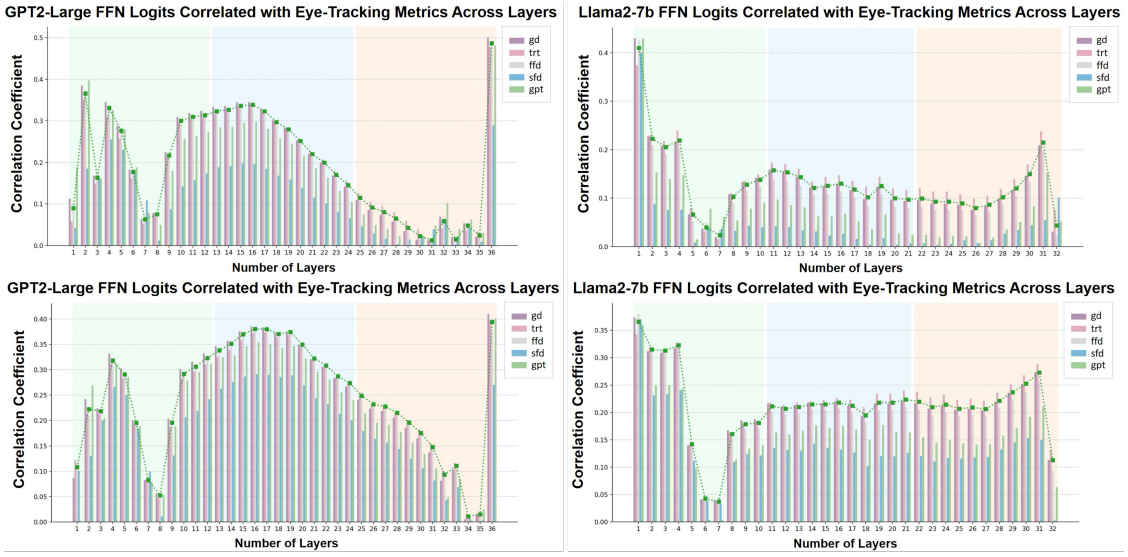


Figure 7.2: Correlation Results Comparison: *Natural Reading (NR) vs. Task-Specific Reading (TSR)*. Correlation results are shown for GPT-2 Large NR (36 layers, top-left), GPT-2 Large TSR (36 layers, bottom-left), Llama2-7B NR (32 layers, top-right), and Llama2-7B TSR (36 layers, bottom-right). Green, blue, and orange boxes indicate the premature, middle, and mature buckets, respectively.

The results indicate that the hidden states of different LLM layers exhibit a clear and strong correlation with human gaze, peaking in the middle bucket and reaching a secondary peak in the mature bucket. This trend is consistent across different eye movement measures and LLMs with varying layer sizes. Considering the nature of these eye movement measures, the increase in correlation in the *premature bucket* suggests that LLMs begin processing tokens by integrating syntactic and semantic features, reflecting an initial focus on token processing. In the *middle bucket*, the further increase in correlation signifies deeper syntactic and semantic processing, with the peak indicating the integration of linguistic features. In the *mature bucket*, the secondary peak likely reflects the final integration of information for word prediction.

The overall trend across the three buckets is similar for GPT-2 and Llama2-7B.

Finding 1. (Layer-wise functionality)

LLM hidden states exhibit a strong correlation with human gaze, characterized by three distinct rises across layers. This pattern suggests a hierarchical progression from initial syntactic and semantic processing to deeper integration and final prediction.

Results and Finding 2. Empirically, LLMs are trained with the next-token prediction objective. Modern LLMs demonstrate strong language understanding and reasoning abilities, raising the question: *Are LLMs merely next-token predictors, or are they task reasoners?* Figure 7.2 presents a comparison between natural reading and task-specific reading. The correlation patterns suggest that LLMs function as both next-token predictors and reasoners, as the trends in task-specific reading closely resemble those observed in natural reading.

Notably, for both GPT-2 and Llama2-7B models, the correlation values in the *middle bucket* during task-specific reading are higher than those observed during natural reading. In particular, for the Llama2-7B model, these values remain consistently higher in task-specific reading. We hypothesize that this indicates Llama2-7B is better suited for reasoning tasks and that the layers in the *middle bucket* and *mature bucket* are activated when processing complex tasks, as it is trained on a larger text corpus and incorporates more advanced post-training techniques.

Finding 2. (LLM functions as both next-token predictor and task reasoner)

LLMs function as both next-token predictors and reasoners, with overall correlation trends aligning with human cognition indicators. Advanced training methods enhance their ability to reason and handle complex tasks.

7.5 Method

7.5.1 Heuristic Steering Layer Selection

A better understanding of LLM mechanisms will help in precisely and efficiently controlling their behaviors, particularly for semantic steering. We argue that the predominant parameter-efficient fine-tuning (PEFT) methods (Han et al., 2024), which by default intervene in the last layer or across all layers, are not optimal[2]. Instead, we propose an efficient heuristic steering layer selection strategy for intervention, based on our cognition-inspired interpretability analysis detailed in Section 7.4.

For semantic steering in LLMs, the layers in the middle bucket are the most suitable candidates for intervention. These layers handle further token processing, information integration, and preliminary reasoning. Additionally, the residual connections (K He et al., 2016) in transformer layers allow the semantic intervention to flow and evolve gradually, avoiding abrupt changes in the final prediction (Chuang et al., 2024).

From a task-oriented perspective, we apply PEFT methods and inference-only methods to the candidate layers in the middle bucket, using a small portion of data, like the validation set, to search for and select the best-performing layer that suits the task scenario. Formally, given M as the best layer for intervention, J represents a set of

2. Recent work (T Yu et al., 2023) on fine-tuning speech translation models also supports our hypothesis.

candidate layers in the middle bucket $\left(\frac{N}{3} \leq M \leq \frac{2N}{3}\right)$. D denotes the validation set. We search for the layer M' that yields the best task score or loss performance, as follows:

$$M' = \arg\max_{l \in M} Score\left(D; P\left(\cdot \mid x_t, l\right)\right). \tag{7.2}$$

Later, we will demonstrate the effectiveness of the heuristic steering layer selection approach in language understanding, reasoning, and generation tasks, as discussed in Section 7.6.

7.5.2 Layer Intervention via Fine-tuning

PEFT approaches, such as additive fine-tuning (i.e., adapters) and reparameterized fine-tuning (i.e., LoRA), are among the most popular due to their efficiency, as they require only a small set of new parameters for task-specific fine-tuning. Let the parameters of an LLM consist of a set of pre-trained, frozen parameters $\phi(\cdot)$ and a set of newly introduced parameters in the inserted block $\psi(\cdot)$. Our layer intervention via fine-tuning to steer semantics in LLM and predicts the next token as follows:

$$y(x_t) = \text{softmax}\Big(\text{logit}_{\phi,\psi}\Big(FFN^N_\phi(x_t) \mid F^M_{\phi,\psi}(x_t), y_{<t}\Big)\Big). \tag{7.3}$$

Here, $FFN^N_\phi(x_t)$ represents the hidden state of the FFN in the final layer with frozen parameters $\phi(\cdot)$, used for token prediction over a vocabulary. M denotes the best layer for semantic intervention, determined by Equation 7.2. $F^M_{\phi,\psi}(x_t)$ indicates that the fusion layer in the LLM integrates new parameters $\psi(\cdot)$ from the newly added block, aligned with the frozen parameters $\phi(\cdot)$.

Our cognitive-inspired selective layer intervention method is an adaptive fine-tuning strategy that identifies the best layer for both effective semantic steering and task performance. Moreover, as our method only operates on a single layer rather than all layers, it significantly reduces computational resources and time, while also avoiding catastrophic forgetting (Luo et al., 2025; H Li et al., 2024).

7.5.3 Layer Intervention during Inference

Efficient semantic steering can be achieved via fine-tuning. However, an even more efficient approach is to steer the semantics of LLMs during inference without introducing additional parameters. Motivated by XL Li et al. (2023) and Leong et al. (2023), which contrast outputs from either a less capable model or outputs induced by a negative prompt, we propose an *implicit layer contrastive intervention* method during inference. First, we fine-tune a contrast model that generates either the desired output or the output we aim to mitigate. In our case, to mitigate toxic token generation, we fine-tune a toxic LLM as the contrast model. Unlike XL Li et al. (2023) and Leong et al. (2023), which contrast the outputs explicitly in the last layer, our method operates on the contextualized value vectors derived from the weight matrices K, Q, V of the attention modules within LLMs. We perform this operation on the best layer for intervention as described in Equation 7.2. Formally, our layer intervention during inference finds the semantic steering direction by contrasting the value vectors as follows:

$$\Delta v^M = v_c^M - v_o^M, \tag{7.4}$$

where v_c^M and v_o^M are the contextualized value vectors of the contrast LLM and the original LLM at the best layer M for semantic intervention. The contextualized value vectors (Elhage et al., 2021) are derived as follows:

$$A^{\ell,h} = \varphi\left(\frac{\left(\mathbf{x}^{\ell-1} W_Q^{\ell,h}\right), \left(\mathbf{x}^{\ell-1} W_K^{\ell,h}\right)^T}{\sqrt{d/H}} + M^{\ell,h}\right), \tag{7.5}$$

$$\mathbf{a}^{\ell} = \sum_{h=1}^{H} A^{\ell,h} \left(\mathbf{x}^{\ell-1} W_V^{\ell,h}\right) W_O^{\ell,h} = \sum_{h=1}^{H} \mathbf{v}^{\ell,h} W_O^{\ell,h}. \tag{7.6}$$

Specifically, $\left(\mathbf{x}^{\ell-1} W_V^{\ell,h}\right)$ represents the attention-weighted, context-sensitive value vector for head h. $\mathbf{v}_i^{\ell,h} \in \mathbb{R}^d$ is the contextualized value vector at position i. We then update the value vector in the layer M of the original LLM:

$$v'^M = v_o^M - \lambda_{\text{norm}}^{\alpha} \cdot \Delta v^M. \tag{7.7}$$

Here, $\lambda_{\text{norm}} = 1 + \left\|\Delta v^M\right\|_2$ is a normalization term that adaptively regulates the steering effect, and α is a hyperparameter that further controls the steering strength. Finally, we preserve the updated steering direction and renormalize the adapted value vector to ensure its representation is close to the original vector:

$$v'^M = v'^M \cdot \frac{\left\|v_o^M\right\|_2}{\left\|v'^M\right\|_2}. \tag{7.8}$$

7.6 Experiment

7.6.1 Datasets and Evaluation

Datasets. We evaluate our proposed efficient semantic steering methods using the General Language Understanding Evaluation (GLUE) benchmark (A Wang et al., 2018), applying selective layer intervention. Specifically, we focus on eight GLUE tasks covering *sentiment analysis* (SST-2), *paraphrase identification* (MRPC, QQP), and *natural language inference* (MNLI-M, MNLI-MM, QNLI, RTE, WNLI). Additionally, to assess our methods in the context of language generation, we examine their effectiveness in natural language toxification and detoxification. To train toxic adapters and contrast models, as described in Sections 7.5.2 and 7.5.3, we use the Toxic Comment Classification Challenge Dataset[3], which contains 15,294 annotated toxic comments. We randomly split this dataset into 13,764 comments for fine-tuning and contrast model training, and 1,530 comments for validation, which is used to determine the optimal layer for intervention, as described in Section 7.5.1.

3. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/overview

Evaluation For the evaluation on the GLUE benchmark, we report both validation-set and test-set *F1 scores* for QQP and MRPC, while *accuracy* is used for all other tasks. Additionally, we use the RealToxicityPrompts (RTP) dataset (Gehman et al., 2020). Following prior work (XL Li et al., 2023; Leong et al., 2023), we sample 2,122 toxic prompts. For each toxic prompt in the RTP dataset, we generate 25 continuations and evaluate their toxicity using the Perspective API [4], which assigns a toxicity score to each continuation, with higher scores indicating a greater likelihood of toxicity. Finally, we use the *average maximum toxicity* as our evaluation metric. Specifically, we compare the toxicity scores and detoxification margins obtained by applying our methods to different layers of the models.

Implementation. For training the GPT-2 adapters, we set the learning rate to 5×10^{-4} and train for 5 epochs. For the Llama2-7B adapters, we adopted the default settings provided by LLaMa-Adapter (R Zhang et al., 2024), using a base learning rate of 9×10^{-3}, a weight decay of 0.02, and training for 5 epochs. For Mistral-7B, we similarly employed the Bottleneck Adapters used for GPT-2, training for 5 epochs with a learning rate of 5×10^{-5} and a weight decay of 0.01. In the detoxification task, we applied the implicit layer contrastive intervention approach with $\alpha = 0.4$ across all models. Following previous works (A Liu et al., 2021; Leong et al., 2023), the model generates 25 continuations per prompt using nucleus sampling with $p = 0.9$, with each continuation limited to a maximum of 20 tokens.

7.6.2 Models and Baselines

We evaluate our efficient semantic steering methods using three sizes of GPT-2 models, as discussed in § 7.4, along with Llama2-7B. We select GPT-2 and Llama2-7B because the former represents earlier classical LLMs, while the latter exemplifies modern LLMs. To ensure generalizability, we also evaluate the Mistral-7B (AQ Jiang et al., 2023) model to assess performance across tasks. Since applying semantic intervention to either the last layer or all layers of LLMs is a conventional approach, we use these two methods as baselines for comparison.

7.6.3 Evaluation on GLUE Benchmark

We first present the performance of our proposed selected layer intervention method on the GLUE Benchmark in Table 7.1. It can be observed that by selecting the optimal layer to steer semantics in LLMs for a specific task, all three LLMs achieve comparable or even superior results compared to conventional all-layer intervention while introducing only $1/N$ of the parameters (N is the number of layers in the LLMs). Specifically, Llama2-7B achieves an absolute increase of **+1.8** on average in the test set, while Mistral-7B achieves an absolute increase of **+4.7** on average in the validation set and **+5.9** in the test set. Moreover, GPT2 achieves comparable or better results on 5 out of 8 tasks in GLUE, Llama2 on 4 tasks, and Mistral on 7 tasks.

Additionally, we find that the optimal layer for our proposed method consistently falls within the middle bucket across all LLMs. For GPT2-L, the best-performing layer is $L19$ for most tasks, while for Llama2-7B and Mistral-7B, it is $L14$ and $L12$, respectively. This aligns with our interpretability analysis and findings based on eye movement

4. https://perspectiveapi.com/

Model	VAL-SET								
	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	WNLI	Avg.
GPT2-L	78.0 $_{l\text{-}19}$	79.5 $_{l\text{-}19}$	85.5 $_{l\text{-}20}$	83.3 $_{l\text{-}19}$	80.6 $_{l\text{-}19}$	71.1 $_{l\text{-}19}$	92.9 $_{l\text{-}19}$	53.5 $_{l\text{-}19}$	78.1
	82.1	83.5	83.1	85.2	82.6	70.4	93.6	53.5	79.2
Llama2-7B	86.4 $_{l\text{-}14}$	87.1 $_{l\text{-}14}$	86.5 $_{l\text{-}14}$	89.3 $_{l\text{-}14}$	83.3 $_{l\text{-}14}$	75.5 $_{l\text{-}14}$	95.8 $_{l\text{-}14}$	56.4 $_{l\text{-}19}$	82.5
	89.0	89.3	86.3	91.9	85.6	65.3	96.7	56.3	82.5
Mistral-7B	87.3 $_{l\text{-}12}$	88.1 $_{l\text{-}12}$	86.9 $_{l\text{-}12}$	91.4 $_{l\text{-}14}$	84.6 $_{l\text{-}12}$	80.1 $_{l\text{-}12}$	95.8 $_{l\text{-}14}$	56.3 $_{l\text{-}12}$	83.8
	89.5	89.7	82.2	81.7	78.1	58.9	96.7	56.3	79.1
Model	**TEST-SET**								
	MNLI-M	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	WNLI	Avg.
GPT2-L	79.3 $_{l\text{-}19}$	79.3 $_{l\text{-}19}$	83.0 $_{l\text{-}20}$	84.1 $_{l\text{-}19}$	65.6 $_{l\text{-}19}$	64.6 $_{l\text{-}19}$	92.4 $_{l\text{-}19}$	58.9 $_{l\text{-}19}$	75.8
	82.6	83.0	82.7	85.6	65.6	62.6	93.5	61.6	77.1
Llama2-7B	82.9 $_{l\text{-}14}$	86.3 $_{l\text{-}14}$	83.4 $_{l\text{-}14}$	88.5 $_{l\text{-}14}$	68.8 $_{l\text{-}14}$	74.7 $_{l\text{-}14}$	95.2 $_{l\text{-}14}$	64.4 $_{l\text{-}19}$	80.5
	89.5	88.8	80.5	92.1	71.6	58.2	93.5	55.5	78.7
Mistral-7B	87.1 $_{l\text{-}12}$	87.5 $_{l\text{-}12}$	86.6 $_{l\text{-}12}$	91.7 $_{l\text{-}14}$	70.5 $_{l\text{-}12}$	81.0 $_{l\text{-}12}$	95.9 $_{l\text{-}14}$	65.8 $_{l\text{-}12}$	83.2
	89.7	89.4	80.4	81.3	62.7	52.3	97.3	65.1	77.3

Table 7.1: Evaluation on GLUE Benchmark. MRPC and QQP are reported using F1, while the other tasks are reported using Accuracy. A green box indicates that the single-layer intervention outperforms the full-layer intervention, an orange box denotes comparable performance, and a blue box indicates slightly lower performance.

measures (see § 7.4). Furthermore, the best layer remains consistent across both the validation and test sets, demonstrating the effectiveness of the heuristic steering layer selection approach (§ 7.5.1).

Lastly, from a task perspective, we find that *paraphrase identification* (MRPC, QQP) and *natural language inference* (MNLI, QNLI, RTE, WNLI) achieve the highest average improvement of **+2.0** across all LLMs compared to full-layer fine-tuning (e.g., **+28.7** for Mistral-7B and **+16.5** for Llama2-7B on the RTE test set). This suggests that, given the complex structure of LLMs, fine-tuning all layers for semantic steering does not always yield the best downstream task performance and can result in parameter redundancy, where certain layers become less active in making accurate predictions. Additionally, the conventional approach of fine-tuning only the last layer of an LLM, based on its proximity to the prediction output, is not optimal. Instead, in practice, the intervention layer should first be identified to better steer semantics and improve prediction accuracy for a specific task.

7.6.4 Analysis on Language Toxification

We evaluate our selective layer intervention via fine-tuning using the language toxification task. Figure 7.3 (*green bars in the upper chart*) shows the toxicity score for inserting a toxic adapter into each layer of the LLMs, with the last layer and full layers (*red line*) used as baselines. The best layer M in the middle bucket for semantic intervention is $L23$ for GPT-2 and $L13$ for Llama2-7B, competing with the results from both the last layer (**+4.5%** for GPT-2 & **+23%** for Llama2) and full layers (**+0.9%** for GPT-2 & **+2.8%** for Llama2). This indicates that the conventional intervention is not optimal, demonstrating

the effectiveness of our proposed selective layer intervention approach, which saves considerable computational resources and time for fine-tuning.

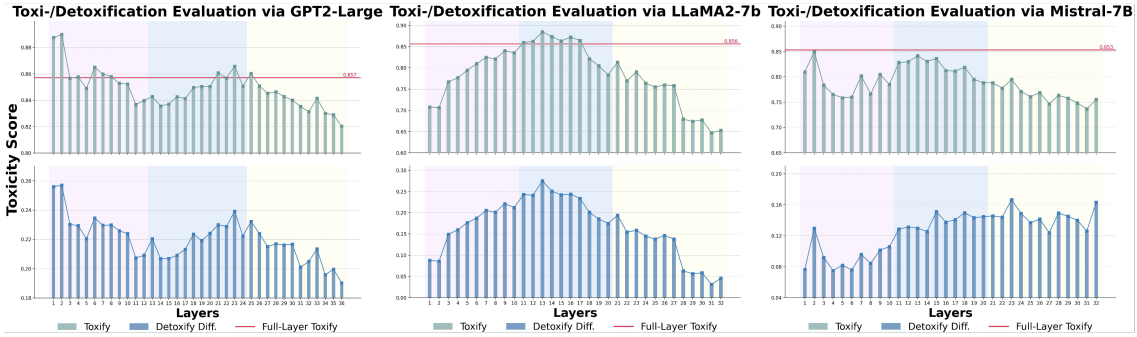


Figure 7.3: Toxification/Detoxification Evaluation. The upper chart (green bars) shows toxicity scores for interventions at each layer via fine-tuning. The bottom chart (blue bars) displays detoxification margin scores per layer during inference, comparing toxicity scores in toxification and detoxification processes. Purple, blue, and orange boxes indicate the premature, middle, and mature buckets.

Furthermore, layer intervention in the mature bucket, including the last layer (used as a baseline), is not remarkable when compared to full-layer intervention for both GPT-2 and Llama2-7B models. According to our behavioral analysis in Section 7.4, the layers in the mature bucket focus more on reasoning and factual knowledge, involving less token processing, which makes them less suitable for semantic steering. Interestingly, for the GPT-2 model, we observe that layers $L1$-$L4$ and $L6$-$L8$ in the premature bucket also outperform the full-layer intervention baseline, whereas this phenomenon is not observed in Llama2-7B. We hypothesize that this difference may be due to the distinct adapter training schemes. For GPT-2, we use the vanilla Adapter as a plug-in, which directly affects token distributions and allows successive minor changes as tokens pass through layers, following the early exit theory (Elbayad et al., 2020; Schuster et al., 2022). However, for Llama2, we use the LLaMa-Adapter (R Zhang et al., 2024), which directly modifies attention. Edits to the attention modules gradually affect token distributions (Geva et al., 2021; Elhage et al., 2021).

We believe that our selective layer intervention approach is LLM-agnostic and not limited to the LLMs used in the interpretability analysis. Figure 7.3 (right) presents the toxicity scores based on Mistral-7B, which align with the observations discussed earlier. Specifically, the best-performing layer in the middle bucket is $L13$, achieving better results than the last layer (**+11.4%**) and nearly identical performance to full-layer intervention (-0.01).

7.6.5 Analysis on Language Detoxification

As a dual task, we evaluate our selective layer intervention during inference using the language detoxification task. This task serves as an adversarial task to mitigate the toxic tokens that are amplified by layer intervention methods during fine-tuning. Thus, we observe the *detoxification margin* in each intervention layer and compare the performance with intervention in the last layer. Figure 7.3 (*blue bars in the bottom chart*) shows the results of the detoxification margin scores. The best layer M for semantic intervention in the middle bucket is $L23$ for GPT-2 and $L16$ for Llama2-7B, both outperforming the last layer results by a large margin (**+2.9%** for GPT and **+24%**

for Llama). Notably, the best layer M for Llama2-7B in the toxification task is $L13$, whereas the best layer for detoxification is $L16$, demonstrating that our heuristic layer selection method is adaptive and suitable for tasks.

Moreover, the detoxification margin scores for layer intervention in the middle bucket for both GPT-2 and Llama2 models are significantly better than those for intervention in the mature bucket. This aligns with our findings that the middle bucket layers are involved in further token processing and information integration, whereas the mature bucket layers focus on reasoning. Finally, for the GPT-2 model, the detoxification margin scores are also significant in the premature bucket layers. This could be attributed to the distinct adapter training scheme, as the earlier layers are more deeply affected by the toxification, which broadens the scope of detoxification. Additional results for GPT-2 small and medium models are reported in Section 7.6.6.

For Mistral-7B, the optimal layer for semantic intervention in the middle bucket is $L15$ for the detoxification task. Although the detoxification margin scores for $L23$ and $L32$ in the mature bucket are slightly higher than that of $L15$ (+0.14 on average), the scores in the middle bucket are more stable. We hypothesize that the differences among GPT-2, Llama2, and Mistral-7B in the mature bucket, where Mistral-7B achieves better detoxification performance, stem from variations in post-training. As an advanced LLM incorporating additional human value alignment for safety, Mistral-7B benefits more from interventions in the mature bucket, leading to improved detoxification performance.

7.6.6 Analysis on Language Toxification and Detoxification on Small Models

For the GPT-2 models, three versions differ in the number of layers. As shown in Figures 7.4 and 7.5, the overall trend in the toxification and detoxification tasks aligns with the observations in Section 7.6. Interestingly, we find that as the number of layers in LLMs increases, the correlation trend and findings become more pronounced.

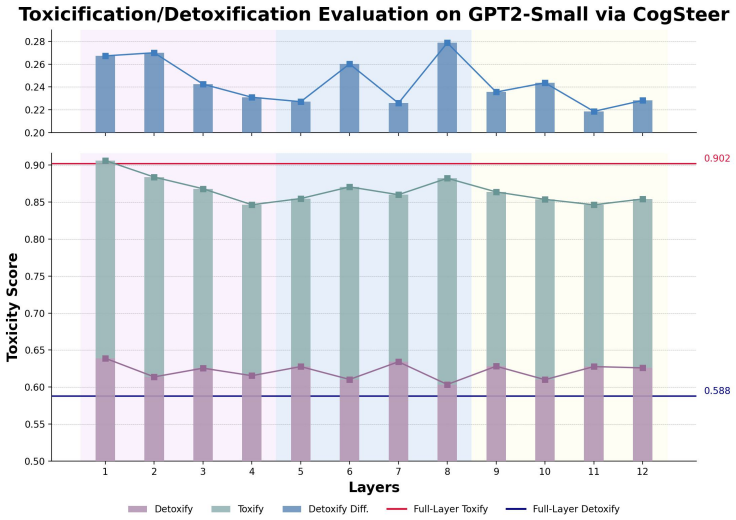


Figure 7.4: Toxification/Detoxification Evaluation on GPT-2 Small (12 Layers). The bottom chart shows toxicity scores for interventions at each layer: green bars represent selective intervention via fine-tuning, and purple bars represent selective intervention during inference. The top chart displays detoxify margin scores per layer, comparing toxicity scores in toxification and detoxification processes. Purple, blue, and orange boxes indicate the premature, middle, and mature buckets.

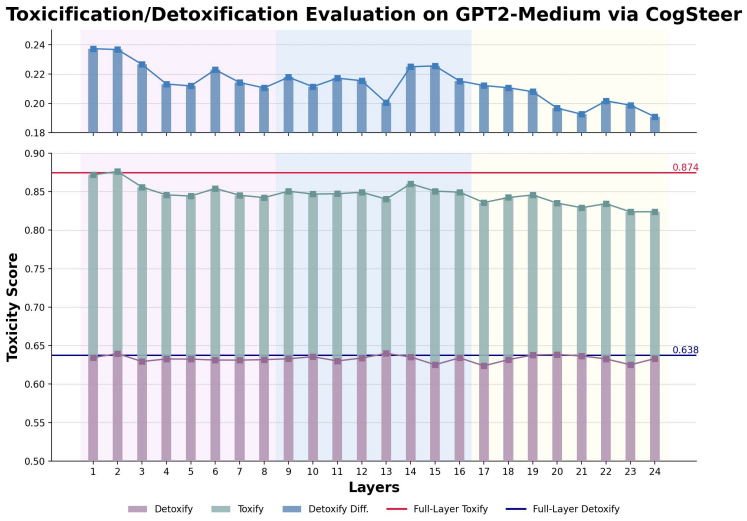


Figure 7.5: Toxification/Detoxification Evaluation on GPT-2 Medium (24 Layers). The bottom chart shows toxicity scores for interventions at each layer: green bars represent selective intervention via fine-tuning, and purple bars represent selective intervention during inference. The top chart displays detoxify margin scores per layer, comparing toxicity scores in toxification and detoxification processes. Purple, blue, and orange boxes indicate the premature, middle, and mature buckets.

7.6.7 Efficiency Analysis

The efficiency analysis of our proposed cognition-inspired selective layer intervention is presented in Table 7.2. By selecting the optimal layer for semantic steering in LLMs, the training time and the number of parameters required for fine-tuning are significantly reduced compared to full-layer intervention. Specifically, our method requires, on average, only **half** the time and **3.0%** of the parameters needed for full-layer settings. Importantly, this reduction in computational cost does not compromise performance; in fact, selective layer intervention performs comparably to, and in some cases even outperforms full-layer intervention.

Model	Time/min	Time%	Params/M	Params%	Toxify Score↑	Detoxify Score↓	Avg. GLUE↑
GPT2-L→ *full*	33	100	14.8	100	0.86	0.60	77.1
→ *single*	13	39.4	0.4	2.7	0.87	0.63	75.8
Llama2-7B→ *full*	205	100	1.3	100	0.86	0.62	78.7
→ *single*	85	41.5	0.04	3.1	0.87	0.59	80.4
Mistral-7B→ *full*	36	100	134.5	100	0.85	0.68	77.3
→ *single*	25	69.4	4.2	3.1	0.84	0.68	83.1

Table 7.2: Efficiency Comparison between *CogSteer (selective single-layer intervention)* and *full-layer intervention*. A green box indicates that the single-layer intervention outperforms the full-layer intervention, while an orange box denotes comparable performance.

7.6.8 Qualitative Analysis

Content warning. Some examples contain harmful or offensive language.

In this section, we conduct a qualitative analysis of the continuations generated under different settings of the GPT2-L model. For language toxification, we compare four continuations produced from a toxic prompt in the Toxic Comment Classification Challenge Dataset (Figure 7.6). The continuation generated by fine-tuning only the last layer is similar to that of the vanilla GPT2-L model. By contrast, fine-tuning on the full model and the selective layer results in more toxic outputs in terms of both toxic words and the overall semantics of the sentence.

For detoxification, we find that the continuation generated using layer contrastive intervention at the last layer still contains offensive words, such as stupid, whereas the output from layer contrastive intervention based on our proposed heuristic steering layer selection is neutral.

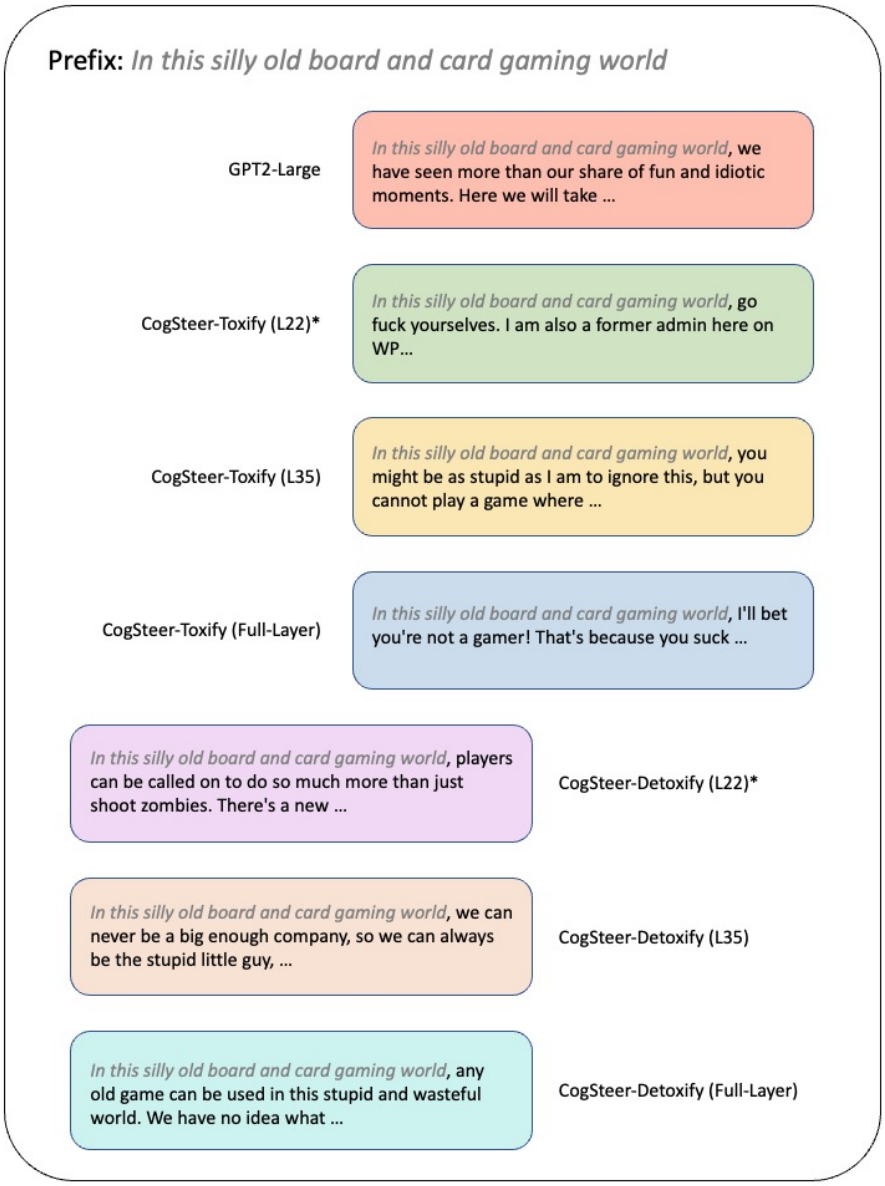


Figure 7.6: Case Study. The continuations are produced from a toxic prompt in the Toxic Comment Classification Challenge Dataset and generated under different settings of the GPT2-L model.

7.7 Conclusion

This study introduces an efficient semantic steering method for LLMs using selective layer intervention. Our approach is motivated by correlation analysis with eye movement measures, making it both interpretable and understandable. Extensive experiments demonstrate that selective layer intervention achieves comparable or even superior performance while significantly reducing training time and the number of parameters required. Overall, our proposed method represents an important step toward improving the interpretability of LLMs and contributes to their safe and efficient deployment.

The art of progress is to preserve order amid change and to preserve change amid order.

— Alfred North Whitehead

8 Conclusion

Contents

8.1 Summary

This dissertation set out to investigate a central question underlying the deployment of foundation models (Bommasani et al., 2021; Brown et al., 2020; Touvron, Lavril, et al., 2023; Achiam et al., 2023; Guo et al., 2025; A Yang et al., 2025): how can systems that demonstrate remarkable general capabilities be made reliably trustworthy in practical use? Rather than treating trustworthiness as a property that can be appended through isolated interventions, the work has approached it as an emergent characteristic of the entire modeling pipeline, shaped by **how data are grounded, how representations are adapted, how outputs are generated, and how behavior can be steered** (Song et al., 2026).

The studies presented in this dissertation collectively show that failures of trustworthiness often arise from disconnections between these stages. Weakly contextualized multimodal supervision can limit the model's ability to anchor meaning in perceptual evidence (Harnad, 1990; Radford et al., 2021; J Li et al., 2022). Adaptation procedures that prioritize efficiency alone may inadvertently disturb pretrained semantic structure (Houlsby et al., 2019; Hu et al., 2022; H Li et al., 2024; Shihab et al., 2026). Generation mechanisms may amplify prior knowledge in ways that override conditioning signals, leading to unfaithful outputs (Rohrbach et al., 2018; Y Li et al., 2023; Hanchao Liu et al., 2024; Bang et al., 2025; Su et al., 2025). Safety interventions that operate only at the surface level risk altering communicative intent (Gehman et al., 2020; Logacheva et al.,

2022; Lee et al., 2024; **Wang**, Liu, et al., 2025), while interpretability analyses that remain descriptive do not necessarily enable actionable control (Belinkov, 2022; Conmy et al., 2023; Rai et al., 2025; **Wang**, Pan, Ding, et al., 2025). Addressing any one of these issues in isolation improves local behavior but does not fully resolve their interaction.

By examining these challenges through complementary empirical studies, the dissertation demonstrates that trustworthiness can be strengthened by reconnecting the stages of the pipeline through targeted, conceptually aligned mechanisms. Context-sensitive grounding provides inputs that more faithfully reflect how multimodal information supports understanding. Stability-aware adaptation preserves the structural knowledge acquired during pretraining while enabling specialization. Inference-time regulation offers a means of constraining generative behavior without requiring continual retraining. Meaning-aware safety reframes alignment as controlled transformation rather than suppression. Finally, cognitively inspired steering links interpretability with operational intervention, enabling efficient and transparent modification of model behavior.

Taken together, these contributions suggest that trustworthy foundation modeling is best understood not as a single technical solution, but as a coordinated design principle: **reliability emerges when grounding, alignment, inference, and control are treated as interdependent elements of a coherent framework**.

8.1.1 Revisiting the Research Questions

Chapter 1 formulated a set of research questions to examine trustworthiness across different stages of the trustworthiness pipeline. The studies presented in Chapters 3– 7 provide corresponding empirical and methodological responses to these questions, allowing them to be revisited in light of the findings.

RQ1 asked whether context-sensitive multimodal resource construction can improve grounded understanding in human-centered settings.

The investigation demonstrates that grounded understanding benefits from data resources explicitly constructed to capture when visual evidence is necessary for interpretation. By developing a context-sensitive multimodal resource that links linguistic complexity, ambiguity, and perceptual relevance, the dissertation shows that grounding can be operationalized as a usage-dependent problem rather than as uniform cross-modal correspondence (Harnad, 1990; Council of Europe, 2001; Lin et al., 2014; **Wang* et al.**, 2022).

RQ2 addressed whether foundation models can be efficiently adapted while preserving the stability of pretrained semantic structure.

The results show that efficient adaptation can be achieved through lightweight, alignment-aware constraints that preserve both inter-modal correspondence and intra-modal structure. By treating adaptation as a constrained transformation of pretrained representations rather than unrestricted task fitting, the dissertation demonstrates that efficiency and stability can be jointly pursued (Houlsby et al., 2019; Hu et al., 2022; Han et al., 2024; H Li et al., 2024; **Wang et al.**, 2023; Shihab et al., 2026).

RQ3 examined whether hallucinations in multimodal generation can be mitigated through inference-time mechanisms rather than additional training.

The findings indicate that hallucinations can be mitigated by intervening in decoding rather than retraining the model. Through inference-time contrastive regulation of competing signals from internal priors and external evidence, the dissertation shows that faithfulness depends not only on learned representations but also on how those

representations are used during generation (XL Li et al., 2023; Leng et al., 2024; Chuang et al., 2024; **Wang**, Pan, Ding, and Biemann, 2024; Su et al., 2025).

RQ4 asked whether safety interventions can remove harmful content while preserving communicative intent and affective meaning.

The work shows that safety can be operationalized as a meaning-aware rewriting problem rather than as simple suppression of undesirable surface forms. By focusing on sentiment-preserving detoxification, the dissertation demonstrates that harmful content can be transformed while retaining communicative intent and affective meaning (Gehman et al., 2020; Dementieva et al., 2025; Lee et al., 2024; **Wang**, Liu, et al., 2025).

RQ5 asked whether human cognitive signals can provide interpretable guidance for identifying and selectively steering internal model behavior.

The analysis demonstrates that human cognitive signals can be converted into actionable guidance for model intervention. By using cognitively motivated, gaze-linked indicators to identify functionally relevant layers and support selective steering, the dissertation connects interpretability to practical controllability and efficiency (Rayner, 1998; Hollenstein et al., 2020; Colman et al., 2022; **Wang**, Li, et al., 2024; **Wang**, Pan, Ding, et al., 2025).

Together, these responses illustrate that trustworthiness can be systematically strengthened when interventions are aligned with the specific stage at which unreliability arises. The research questions are therefore not independent problems, but interconnected facets of a broader effort to align foundation model behavior with grounded inputs, stable representations, faithful generation, and interpretable control.

8.1.2 Key Insights for Trustworthy Foundation Models

Beyond answering the individual research questions, this dissertation yields several broader insights into how trustworthiness should be approached in the design and study of foundation models. These insights extend beyond the specific methods presented and point toward methodological principles for future work in this area.

First, trustworthiness must be enforced across stages rather than appended locally.

Many existing efforts address reliability through isolated modifications, such as adding alignment objectives, filtering outputs, or introducing post-hoc corrections (Ouyang et al., 2022; Rafailov et al., 2023; Y Bai et al., 2022; S Yin et al., 2024; Song et al., 2026). The findings of this dissertation suggest that such localized interventions are most effective when they are coordinated with the stage of the pipeline in which unreliability originates. Grounded data design, stability-aware adaptation, inference-time regulation, and interpretable steering operate on different layers of the system, and their effects are complementary rather than interchangeable. *Trustworthiness therefore emerges from continuity across stages, not from a single corrective mechanism.*

Second, inference-time behavior is as critical as training-time learning.

Traditional perspectives emphasize improving model parameters through larger datasets or more sophisticated optimization. However, generative foundation models remain highly sensitive to how learned knowledge is deployed during decoding. The results presented here show that carefully designed inference-time strategies can meaningfully influence faithfulness and safety without requiring continual retraining (A Liu et al., 2021; XL Li et al., 2023; Sennrich et al., 2024; Chuang et al., 2024; Su et al., 2025; C Li et al., 2026). *This highlights the importance of viewing inference not merely as execution, but as a controllable component of the modeling process.*

Third, efficiency and reliability need not be opposing objectives.

Efficiency is often pursued primarily for scalability and deployment considerations, while reliability is treated as an additional constraint. The studies in this dissertation indicate that structured, lightweight interventions can in some cases enhance both properties simultaneously by preserving the organization of pretrained knowledge and avoiding unnecessary parameter modification (Houlsby et al., 2019; Hu et al., 2022; Poth et al., 2023; Han et al., 2024; Shihab et al., 2026; K Yao et al., 2026). *This suggests that efficiency, when guided by alignment-aware principles, can support rather than undermine trustworthy behavior.*

Fourth, human-centered signals can serve as structural guidance rather than only evaluation criteria.

Human information—whether reflected in perceptual grounding, affective intent, or cognitive processing patterns—is frequently used to assess model performance after the fact. The work presented here demonstrates that such signals can also inform where and how models should be guided, providing actionable structure for both data construction and model intervention (Harnad, 1990; Rayner, 1998; Hollenstein et al., 2020; Colman et al., 2022). *Incorporating these signals shifts the role of human knowledge from external supervision to an integral component of model design.*

Fifth, trustworthiness must be managed as a set of interacting trade-offs rather than a collection of independent objectives.

The studies in this dissertation repeatedly show that improving one aspect of model behavior may place pressure on another. Efficient adaptation can conflict with the preservation of pretrained structure, safety interventions can weaken communicative fidelity, and strong generative priors can override grounded evidence during decoding. A trustworthy model must therefore be designed not simply to optimize isolated properties, but to negotiate these tensions in ways that preserve the overall coherence of system behavior. *This suggests that trustworthiness is inherently multi-objective: progress depends on balancing competing constraints rather than maximizing a single criterion.*

Taken together, these insights point toward a conception of trustworthy foundation modeling as an interdisciplinary endeavor that integrates representation learning, inference control, and human-centered understanding. Rather than viewing reliability as a constraint imposed on otherwise autonomous systems, it can be framed as a design principle embedded throughout the lifecycle of foundation models.

8.2 Limitations

While this dissertation presents a set of approaches for improving the trustworthiness of foundation models, its scope is intentionally bounded, and several limitations should be acknowledged to clarify the context in which the findings apply.

First, the proposed methods operate primarily within existing foundation model architectures. The studies assume widely adopted transformer-based language and vision–language models as their underlying substrates (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2021; J Li et al., 2023). As a result, the contributions focus on how such models can be better grounded, adapted, and steered, rather than proposing fundamentally new architectures. Although this enables practical applicability, it also means that the conclusions are shaped by the representational assumptions and constraints of current model designs.

Second, the integration of human-centered signals is explored in targeted forms rather than as a unified learning paradigm. Cognitive and affective information is used to guide dataset construction, evaluation, or selective intervention, but not yet incorporated as a continuous signal during large-scale pretraining. Extending these signals into earlier stages of model development remains an open challenge requiring both methodological advances and suitable data resources.

Third, evaluation of trustworthiness remains task-dependent and lacks universally accepted metrics. Groundedness, faithfulness, and meaning-preserving safety are assessed through combinations of quantitative benchmarks and structured analyses, reflecting current practice in the field (Liang et al., 2023; Fu et al., 2025; Y Li et al., 2023; Bang et al., 2025; Dementieva et al., 2025; Song et al., 2026). However, standardized evaluation frameworks that jointly measure these dimensions are still evolving. Consequently, the results should be interpreted as evidence across complementary tasks rather than as a single unified score of trustworthiness.

Fourth, the dissertation advances a unified perspective on trustworthy foundation models at the conceptual and methodological levels, but it does not present a single end-to-end system that integrates all proposed components within one shared architecture or evaluation pipeline. The contributions instead consist of complementary studies that intervene at different stages of the model lifecycle, each targeting a distinct source of unreliability. This design makes it possible to analyze specific mechanisms with clarity, but it also means that the dissertation does not empirically validate a fully integrated framework in which grounding, adaptation, inference control, safety transformation, and cognitively informed steering are optimized jointly.

Fifth, the focus of this dissertation is on functional reliability rather than long-term interactional or societal dynamics. The work examines how models behave under controlled conditions of grounding, adaptation, and generation, but does not address broader questions such as longitudinal user interaction, cultural variability in interpretation, or policy-level considerations (Bender et al., 2021; Weidinger et al., 2021; Cao et al., 2024). These aspects are important for deployment but extend beyond the technical scope considered here.

These limitations do not diminish the contributions of the dissertation; rather, they delineate the boundary between what can currently be achieved through model- and data-centric interventions and what remains to be explored in future research. Recognizing these boundaries helps situate the present work as a step toward more comprehensive frameworks for trustworthy foundation modeling.

8.3 Future Work

The findings of this dissertation point toward several avenues for further research that extend the goal of trustworthy foundation modeling beyond the specific methods explored here. These directions are not tied to particular implementations, but rather reflect broader conceptual developments suggested by the results.

First, *future work may investigate how human-centered signals can be incorporated earlier in the learning process.* The present studies demonstrate the value of perceptual grounding, affective alignment, and cognitively inspired guidance at stages such as dataset design and model intervention. An important next step is to examine how such signals might inform representation learning itself, allowing models to internalize

context sensitivity and interpretability during pretraining rather than relying solely on post hoc adjustments (Harnad, 1990; Radford et al., 2021; Hollenstein et al., 2020).

Second, *there is a need for more unified frameworks for inference-time control.* This dissertation highlights the role of decoding and intervention mechanisms in shaping model behavior, suggesting that inference should be treated as an adaptable component rather than a fixed execution phase. Recent reasoning-oriented and multimodal systems further suggest that inference is becoming a resource-adaptive process, where models allocate computation, retrieval, uncertainty estimation, or deliberation according to task demands (Hurst et al., 2024; Jaech et al., 2024; Guo et al., 2025; A Yang et al., 2025; DeepMind, 2025; C Li et al., 2026). Developing general principles for controllable generation—capable of balancing faithfulness, safety, and flexibility across tasks—could provide a systematic foundation for managing model outputs in diverse applications (A Liu et al., 2021; XL Li et al., 2023; Lewis et al., 2020; Y Bai et al., 2022; Rafailov et al., 2023; Su et al., 2025).

Third, *evaluation methodologies must evolve to reflect the multidimensional nature of trustworthiness.* Current benchmarks often assess individual properties, such as accuracy or toxicity reduction, in isolation. The work presented here indicates that reliability depends on the interaction of grounding, stability, faithfulness, and controllability. Designing evaluation protocols that jointly measure these dimensions, potentially incorporating human-centered criteria alongside automated metrics, will be essential for capturing real-world model behavior more faithfully (Liang et al., 2023; Fu et al., 2025; Y Li et al., 2023; Hanchao Liu et al., 2024; Bang et al., 2025; Dementieva et al., 2025; Song et al., 2026).

Finally, *advancing trustworthy foundation models will require closer integration between technical development and insights from human cognition and communication.* Understanding how people interpret language, attend to information, and maintain intent in interaction can inform the design of systems that behave in ways that are not only statistically effective but also contextually appropriate and interpretable (Rayner, 1998; Hollenstein et al., 2020; Colman et al., 2022; Cao et al., 2024). Such interdisciplinary perspectives may help bridge the remaining gap between high-performing models and systems that can be reliably embedded in human environments.

In summary, the path forward lies in treating trustworthiness not as a constraint added to existing models, but as a guiding principle for how future foundation models are trained, adapted, evaluated, and controlled. Advancing this goal will require keeping *vision* closely tied to grounded evidence, *language* closely tied to faithful and meaning-preserving generation, and *gaze* and other human cognitive signals closely tied to interpretable intervention. By continuing to connect these perspectives, research can move toward AI systems whose capabilities are matched by their dependability.

8.4 Ethics Statement

This dissertation investigates methods for improving the trustworthiness of foundation models through multimodal grounding, controlled generation, and interpretable model steering. The research draws on publicly available resources together with data constructed under established annotation protocols, and it does not involve the acquisition of personal or sensitive user information.

Data Sources. All datasets used in this work originate from publicly accessible corpora or benchmarks with established research usage conditions. The MOTIF dataset is constructed from the Wikipedia Simple Corpus together with images drawn from the MS COCO 2017 collection (Benzahra and Yvon, 2019; Lin et al., 2014). The representation-learning study in Chapter 4 employs standard multimodal benchmarks, including MS COCO and Flickr30K, for cross-modal alignment and retrieval evaluation (Lin et al., 2014; Plummer et al., 2015). The hallucination analysis presented in Chapter 5 relies exclusively on publicly released evaluation resources, such as POPE, MME, LLaVA-Bench, and related vision–language benchmarks (Y Li et al., 2023; Fu et al., 2025; Haotian Liu et al., 2023), all of which are derived from existing datasets including MS COCO and other open evaluation suites. The sentiment-consistent rewriting dataset introduced in Chapter 6 is built from publicly available Chinese-language text sources that were filtered and rewritten through controlled annotation procedures to remove unsuitable or sensitive material (Junyu Lu et al., 2023; Deng et al., 2022; J Zhou et al., 2022; A Jiang et al., 2022; Q Yang et al., 2025). Finally, the cognitive analysis in Chapter 7 uses established eye-tracking corpora, including ZuCo 2.0, GECO, and Provo, which contain anonymized experimental data collected in prior controlled studies (Hollenstein et al., 2020; Colman et al., 2022; Luke and Christianson, 2018). No new behavioral or biometric human-subject experiments were conducted as part of this dissertation. Newly collected human input was limited to structured annotation and validation tasks carried out under the procedures described below.

Annotation and Data Construction. Where manual annotation was required, such as in the construction of context-sensitive multimodal examples and sentiment-consistent rewriting data, tasks were carried out through structured crowd-based or expert validation workflows. Annotators were provided with explicit task descriptions and quality guidelines, and no personally identifying information was collected. Annotation focused solely on linguistic interpretation, contextual relevance, and transformation quality.

Handling of Potentially Harmful Content. Research on toxicity mitigation necessarily involves the examination of offensive language (Gehman et al., 2020; Logacheva et al., 2022; Dementieva et al., 2025). In this work, such material is included only to develop methods that reduce harmful model behavior while preserving communicative intent. Data instances deemed inappropriate for constructive rewriting (e.g., targeting protected groups or involving explicit abuse) were excluded during dataset curation. The objective is not to reproduce harmful content, but to study how models can transform it into safer alternatives.

Prompt Transparency and Reproducibility. All prompting strategies used for evaluation or data construction are explicitly documented in the corresponding chapters. Prompt templates and processing procedures are released alongside the implementation to ensure transparency and reproducibility of experimental results.

Open Research Artifacts. Code, processing scripts, and derived resources developed in this dissertation are publicly released to support verification and reuse by the research community. These materials are intended solely for scientific research and must be used in accordance with the licenses of the underlying datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. "GPT-4 technical report," arXiv: 2303.08774. (Cited on pages 1, 16, 22, 67, 124).

Meta AI. 2025. *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.* https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-05-20. (Cited on page 95).

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, and Malcolm Reynolds. 2022. *Flamingo: A Visual Language Model for Few-Shot Learning*. In *Advances in Neural Information Processing Systems,* 35:1–21. New Orleans, Louisiana, USA. (Cited on pages 18 sqq., 55, 69).

Mohammed H. Albahiri and Ali Albashir Mohammed Alhaj. 2020. "Role of visual element in spoken English discourse: implications for YouTube technology in EFL classrooms." *The Electronic Library* 38 (3): 531–544. (Cited on page 37).

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. "Layer Normalization," arXiv: 1607.06450. (Cited on page 16).

Nikolay Babakov, David Dale, Varvara Logacheva, and Alexander Panchenko. 2022. *A large-scale computational study of content preservation measures for text style transfer and paraphrase generation*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop,* 300–321. Dublin, Ireland: Association for Computational Linguistics. (Cited on page 33).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. "Neural Machine Translation by Jointly Learning to Align and Translate," arXiv: 1409.0473. (Cited on page 26).

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond," arXiv: 2308.12966. (Cited on page 69).

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. "Qwen3-VL Technical Report," arXiv: 2511.21631. (Cited on page 20).

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. "Qwen2.5-VL Technical Report," arXiv: 2502.13923. (Cited on pages 20 sq.).

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, and Cameron McKinnon. 2022. "Constitutional AI: Harmlessness from AI Feedback," arXiv: 2212.08073. (Cited on pages 126, 129).

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. *HalluLens: LLM Hallucination Benchmark*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 24128–24156. Vienna, Austria: Association for Computational Linguistics. (Cited on pages 3 sq., 32, 124, 128 sq.).

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. 2022. "VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts." *Advances in Neural Information Processing Systems* 35. (Cited on pages 18, 32, 52, 55 sq., 69).

Yonatan Belinkov. 2022. "Probing Classifiers: Promises, Shortcomings, and Advances." *Computational Linguistics* 48 (1): 207–219. (Cited on pages 34, 109, 125).

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610–623. Online: Association for Computing Machinery. (Cited on pages 2, 33, 128).

Marc Benzahra and François Yvon. 2019. *Measuring text readability with machine comprehension: a pilot study*. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications,* 412–422. Florence, Italy: Association for Computational Linguistics. (Cited on pages 43, 49, 130).

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. "Language models can explain neurons in language models." *OpenAI Blog,* (cited on page 111).

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, and Emma Brunskill. 2021. "On the Opportunities and Risks of Foundation Models," arXiv: 2108.07258. (Cited on pages 1, 14, 22, 124).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33. (Cited on pages 1, 16, 22, 32, 55, 124).

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. "Concreteness Ratings for 40 Thousand Generally Known English Word Lemmas." *Behavior research methods* 46 (3): 904–911. (Cited on page 44).

Yang Trista Cao, Lovely-Frances Domingo, Sarah Gilbert, Michelle L. Mazurek, Katie Shilton, and Hal Daumé III. 2024. *Toxicity Detection is NOT all you Need: Measuring the Gaps to Supporting Volunteer Content Moderators through a User-Centric Method*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 3567–3587. Miami, Florida, USA: Association for Computational Linguistics, November. (Cited on pages 84, 128 sq.).

Xiaobin Chen and Detmar Meurers. 2016. *Characterizing Text Difficulty with Word Frequencies*. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications,* 84–94. San Diego, California, USA: Association for Computational Linguistics. (Cited on page 49).

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *UNITER: UNiversal Image-TExt Representation Learning*. In *Computer Vision – European Conference on Computer Vision (ECCV) 2020,* 12375:104–120. Lecture Notes in Computer Science. Glasgow, UK: Springer. (Cited on pages 18, 30, 40, 55 sq., 61).

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, and Joseph E. Gonzalez. 2023. "Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality." *LMSYS Organization,* (cited on page 70).

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. *DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models*. In *International Conference on Learning Representations,* 1–26. Vienna, Austria: OpenReview.net. (Cited on pages 26, 32 sq., 112 sqq., 126).

Toon Colman, Margot Fonteyne, Joke Daems, Nicolas Dirix, and Lieve Macken. 2022. *GECO-MT: The Ghent Eye-tracking Corpus of Machine Translation*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 29–38. Marseille, France: European Language Resources Association. (Cited on pages 29, 34, 110, 112, 126 sq., 129 sq.).

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. *Towards Automated Circuit Discovery for Mechanistic Interpretability*. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023),* 16318–16352. New Orleans, Louisiana, USA. (Cited on pages 34, 109, 111, 125).

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge*. U.K: Press Syndicate of the University of Cambridge. (Cited on pages 38, 40, 125).

Luise D.ürlich and Thomas François. 2018. *EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* 1–7. Miyazaki, Japan: European Language Resources Association (ELRA). (Cited on page 44).

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. *Knowledge Neurons in Pretrained Transformers*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 8493–8502. Dublin, Ireland: Association for Computational Linguistics. (Cited on pages 17, 34, 109, 111).

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. *InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning*. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023),* 1–25. New Orleans, Louisiana, USA. (Cited on page 2).

Bridget Dalton and Dana L. Grisham. 2011. "eVoc Strategies: 10 Ways to Use Technology to Build Vocabulary." *Reading Teacher* 64 (5): 306–317. (Cited on page 37).

Google DeepMind. 2025. *Gemini 2.5 Pro: Our Most Intelligent AI Model.* https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-05-20. (Cited on pages 86, 129).

DeepSeek-AI. 2024. *DeepSeek-V3 Technical Report*. arXiv: 2412.19437. (Cited on page 86).

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. 2023. *Scaling Vision Transformers to 22 Billion Parameters*. In *Proceedings of the 40th International Conference on Machine Learning,* 202:7480–7512. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: PMLR. (Cited on pages 52, 54).

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, **Xintong Wang**, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. *Multilingual and Explainable Text Detoxification with Parallel Corpora.* In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025),* 7998–8025. Abu Dhabi, UAE: Association for Computational Linguistics. (Cited on pages 33, 40, 87, 111, 126, 128 sqq.).

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, **Xintong Wang**, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. *Overview of the Multilingual Text Detoxification Task at PAN 2024*. In *Proceedings of the 2024 Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024),* 2432–2461. Grenoble, France. (Cited on page 87).

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. *COLD: A Benchmark for Chinese Offensive Language Detection*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 11580–11599. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 86 sq., 130).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171–4186. Minneapolis, Minnesota, USA: Association for Computational Linguistics. (Cited on pages 14, 16, 52, 55 sq., 127).

Haiwen Diao, Ying Zhang, Wei Liu, Xiang Ruan, and Huchuan Lu. 2023. "Plug-and-Play Regulators for Image-Text Matching." *IEEE Transactions on Image Processing,* (cited on pages 52, 55, 61).

Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. 2023. "Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation." *Information Fusion* 99:101896. (Cited on page 2).

Xiangjue Dong, Maria Teleki, and James Caverlee. 2024. "A Survey on LLM Inference-Time Self-Improvement," arXiv: 2412.14352. (Cited on page 110).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. In *Ninth International Conference on Learning Representations (ICLR 2021),* 1–22. Online: OpenReview.net. (Cited on pages 18, 40, 52, 54, 56, 61).

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and Angela Fan. 2024. "The Llama 3 Herd of Models," arXiv: 2407.21783. (Cited on page 109).

Jean Ecalle, Annie Magnan, Houria Bouchafa, and Jean Emile Gombert. 2009. "Computer-Based Training with Ortho-Phonological Units in Dyslexic Children: New Investigations." *Dyslexia* 15 (3): 218–238. (Cited on page 37).

Grant Eckstein, Wesley Schramm, Madeline Noxon, and Jenna Snyder. 2019. "Reading L1 and L2 Writing: An Eye-tracking Study of TESOL Rater Behavior." *TESL-EJ* 23 (1). (Cited on page 29).

Maha Elbayad, Jiatao Gu, Edouard Grave, and Michael Auli. 2020. *Depth-adaptive Transformer*. In *Eighth International Conference on Learning Representations (ICLR 2020),* 1–15. Online: OpenReview.net. (Cited on page 119).

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and Tom Conerly. 2021. *A mathematical framework for transformer circuits. Transformer Circuits Thread*. In *Transformer Circuits Thread.* (Cited on pages 34, 116, 119).

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. *VSE++: Improving Visual-Semantic Embeddings with Hard Negatives*. In *Proceedings of the British Machine Vision Conference (BMVC),* 1–13. Newcastle upon Tyne, UK: BMVA Press. (Cited on pages 18 sq., 30, 52, 56).

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. *Strong Baselines for Complex Word Identification across Multiple Languages*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 970–977. Minneapolis, Minnesota, USA: Association for Computational Linguistics. (Cited on page 40).

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, and Xing Sun. 2025. *MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models*. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025),* 1–12. San Diego, California, USA. (Cited on pages 32, 68, 128 sqq.).

Changjiang Gao, Shujian Huang, Jixing Li, and Jiajun Chen. 2023. *Roles of Scaling and Instruction Tuning in Language Perception: Model vs. Human Attention*. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 13042–13055. Singapore: Association for Computational Linguistics. (Cited on page 111).

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. *RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models*. In *Findings of the Association for Computational Linguistics: EMNLP 2020,* 3356–3369. Online: Association for Computational Linguistics. (Cited on pages 3, 5, 33, 117, 124, 126, 130).

Emilie Gerbier, G.érard Bailly, and Marie L. Bosse. 2018. "Audio–visual Synchronization in Reading while Listening to Texts: Effects on Visual Behavior and Verbal Learning." *Computer Speech & Language* 47:74–92. (Cited on page 37).

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. *Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space*. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing,* 30–45. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 34, 109, 111).

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. *Transformer Feed-Forward Layers Are Key-Value Memories*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP),* 5484–5495. Online: Association for Computational Linguistics. (Cited on pages 17, 34, 112 sq., 119).

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. *Patchscopes: A Unifying Framework for Inspecting Hidden Representations of Language Models*. In *Proceedings of the 41st International Conference on Machine Learning,* 235:15466–15490. Proceedings of Machine Learning Research. Vienna, Austria: PMLR. (Cited on pages 34, 109).

Ross Girshick. 2015. *Fast R-CNN*. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 1440–1448. Santiago, Chile: IEEE. (Cited on pages 18, 52, 55 sq.).

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. "Localizing Model Behavior with Path Patching," arXiv: 2304.05969. (Cited on page 111).

Sian Gooding and Ekaterina Kochmar. 2018. *CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting*. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications,* 184–194. New Orleans, Louisiana, USA: Association for Computational Linguistics. (Cited on page 40).

Sian Gooding and Ekaterina Kochmar. 2019. *Complex Word Identification as A Sequence Labelling Task*. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 1148–1153. Florence, Italy: Association for Computational Linguistics. (Cited on page 40).

Marco Guerini, Lorenzo Gatti, and Marco Turchi. 2013. *Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet*. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* 1259–1269. Seattle, Washington, USA: Association for Computational Linguistics. (Cited on page 33).

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. "Detecting and Preventing Hallucinations in Large Vision Language Models." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (16): 18135–18143. (Cited on pages 32, 68 sq.).

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, and Xiao Bi. 2025. "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," arXiv: 2501.12948. (Cited on pages 109, 124, 129).

Noemi Hahn, John J. Foxe, and Sophie Molholm. 2014. "Impairments of Multisensory Integration and Cross-sensory Learning as Pathways to Dyslexia." *Neuroscience & Biobehavioral Reviews* 47:384–392. (Cited on page 37).

Zeyu Han, Chao Gao, Jinyang Liu, and Sai Qian Zhang. 2024. "Parameter-efficient fine-tuning for large models: A comprehensive survey," arXiv: 2403.14608. (Cited on pages 32, 111, 114, 125, 127).

Stevan Harnad. 1990. "The Symbol Grounding Problem." *Physica D: Nonlinear Phenomena* 42 (1–3): 335–346. (Cited on pages 4, 31, 124 sq., 127, 129).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. *Deep residual learning for image recognition*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 770–778. Las Vegas, Nevada, USA: IEEE. (Cited on pages 56, 59, 61, 114).

Shwai He, Liang Ding, Daize Dong, Jeremy Zhang, and Dacheng Tao. 2022. *SparseAdapter: An Easy Approach for Improving the Parameter-Efficiency of Adapters*. In *Findings of the Association for Computational Linguistics: EMNLP 2022,* 2184–2190. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on page 23).

Jonathan Ho and Tim Salimans. 2021. *Classifier-Free Diffusion Guidance*. In *Deep Generative Models and Downstream Applications Workshop at NeurIPS 2021,* 1–14. Online: OpenReview.net. (Cited on page 75).

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. *ZuCo 2.0: A Dataset of Physiological Recordings During Natural Reading and Annotation*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC),* 138–146. Marseille, France: European Language Resources Association. (Cited on pages 29, 34, 110, 112, 126 sq., 129 sq.).

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. *The Curious Case of Neural Text Degeneration*. In *International Conference on Learning Representations,* 1–16. Online. (Cited on page 26).

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. *Parameter-Efficient Transfer Learning for NLP*. In *Proceedings of the 36th International Conference on Machine Learning (ICML),* 2790–2799. Long Beach, California, USA: PMLR. (Cited on pages 4, 23 sq., 31, 53, 55, 59, 111, 124 sq., 127).

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. *LoRA: Low-Rank Adaptation of Large Language Models*. In *Tenth International Conference on Learning Representations (ICLR 2022),* 1–26. Online: OpenReview.net. (Cited on pages 4, 24 sq., 31, 111, 124 sq., 127).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, and Bing Qin. 2023. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," arXiv: 2311.05232. (Cited on page 2).

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. "Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers," arXiv: 2004.00849. (Cited on page 61).

Drew A. Hudson and Christopher D. Manning. 2019. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 6700–6709. Long Beach, California, USA: IEEE. (Cited on pages 4, 30 sq., 74).

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, and Alec Radford. 2024. "GPT-4o system card," arXiv: 2410.21276. (Cited on pages 86, 129).

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. *Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI*. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21),* 624–635. Online: Association for Computing Machinery. (Cited on page 2).

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, and Alex Carney. 2024. "OpenAI o1 system card," arXiv: 2412.16720. (Cited on pages 86, 129).

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. *Visual prompt tuning*. In *Computer Vision – European Conference on Computer Vision (ECCV) 2022,* 13693:709–727. Lecture Notes in Computer Science. Tel Aviv, Israel: Springer. (Cited on pages 23, 31, 55).

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. "SWSR: A Chinese dataset and lexicon for online sexism detection." *Online Social Networks and Media* 27:100182. (Cited on pages 86 sq., 130).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and Lucile Saulnier. 2023. *Mistral 7B*. arXiv: 2310.06825. (Cited on page 117).

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. "Hallucination Augmented Contrastive Learning for Multimodal Large Language Model." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 27036–27046. (Cited on pages 68 sqq.).

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. *FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models*. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 5042–5063. Miami, Florida, USA: Association for Computational Linguistics. (Cited on page 69).

Andrej Karpathy and Li Fei-Fei. 2015. *Deep Visual-Semantic Alignments for Generating Image Descriptions*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 3128–3137. Boston, Massachusetts, USA: IEEE. (Cited on pages 18, 30, 60).

Md Tawkat Islam Khondaker, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. 2024. "DetoxLLM: A Framework for Detoxification with Explanations," arXiv: 2402.15951. (Cited on page 33).

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision*. In *Proceedings of the 38th International Conference on Machine Learning,* 139:5583–5594. Online: PMLR. (Cited on pages 18, 55, 61).

J. Peter Kincaid, James A. Aagard, and John W. O.'hara. 1980. *Development and Test of a Computer Readability Editing System (CRES)*. Technical report TAEG-R-83. Orlando, Florida, USA: Naval Training Analysis and Evaluation Group. (Cited on page 43).

Ekaterina Kochmar, Sian Gooding, and Matthew Shardlow. 2020. *Detecting Multiword Expression Type Helps Lexical Complexity Assessment*. In *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 4426–4435. Marseille, France: European Language Resources Association. (Cited on page 50).

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, and David A. Shamma. 2017. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." *International Journal of Computer Vision* 123 (1): 32–73. (Cited on pages 31, 56, 61).

Beomseok Lee, Hyunwoo Kim, Keon Kim, and Yong Suk Choi. 2024. *XDetox: Text Detoxification with Token-Level Toxicity Explanations*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 15215–15226. Miami, Florida, USA: Association for Computational Linguistics, November. (Cited on pages 33, 94, 125 sq.).

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. *Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 13872–13882. Seattle, Washington, USA: IEEE. (Cited on pages 5, 26, 32 sq., 69, 74 sqq., 126).

Chak Tou Leong, Yi Cheng, Jiashuo Wang, Jian Wang, and Wenjie Li. 2023. *Self-Detoxifying Language Models via Toxification Reversal*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 4433–4449. Singapore: Association for Computational Linguistics. (Cited on pages 33, 111, 115, 117).

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. *The Power of Scale for Parameter-Efficient Prompt Tuning*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,* 3045–3059. Online: Association for Computational Linguistics. (Cited on page 22).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K.üttler, Mike Lewis, Wen-tau Yih, and Tim Rocktäschel. 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020),* 9459–9474. Online. (Cited on page 129).

Changcheng Li, Jiancan Wu, Hengheng Zhang, Zhengsu Chen, Guo An, Junxiang Qiu, Xiang Wang, and Qi Tian. 2026. "Confidence Before Answering: A Paradigm Shift for Efficient LLM Uncertainty Estimation," arXiv: 2603.05881. (Cited on pages 5, 33, 126, 129).

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. *Revisiting Catastrophic Forgetting in Large Language Model Tuning*. In *Findings of the Association for Computational Linguistics: EMNLP 2024,* 4297–4308. Miami, Florida, USA: Association for Computational Linguistics. (Cited on pages 4, 22, 24, 32, 115, 124 sq.).

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. In *Proceedings of the 40th International Conference on Machine Learning,* 202:19730–19742. Proceedings of Machine Learning Research. Honolulu, Hawaii, USA: PMLR. (Cited on pages 2, 67, 127).

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. In *Proceedings of the 39th International Conference on Machine Learning,* 162:12888–12900. Baltimore, Maryland, USA: PMLR. (Cited on pages 2, 17 sqq., 22, 30, 52, 55, 64, 69, 124).

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation." *Advances in Neural Information Processing Systems* 34. (Cited on pages 18, 31 sq., 52, 55).

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. *Contrastive Decoding: Open-ended Text Generation as Optimization*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 12286–12312. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 5, 26 sq., 32 sq., 73, 75, 79, 112, 115, 117, 126, 129).

Xiang Lisa Li and Percy Liang. 2021. *Prefix-Tuning: Optimizing Continuous Prompts for Generation*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 4582–4597. Online: Association for Computational Linguistics. (Cited on pages 22 sq.).

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, and Furu Wei. 2020. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*. In *Computer Vision – European Conference on Computer Vision (ECCV) 2020,* 12375:121–137. Lecture Notes in Computer Science. Glasgow, UK: Springer. (Cited on pages 18, 30, 40, 55).

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. *Evaluating Object Hallucination in Large Vision-Language Models*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 292–305. Singapore: Association for Computational Linguistics. (Cited on pages 3 sq., 32, 68, 124, 128 sqq.).

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, and Ananya Kumar. 2023. "Holistic Evaluation of Language Models." *Transactions on Machine Learning Research,* (cited on pages 2, 128 sq.).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. *Microsoft COCO: Common Objects in Context*. In *Computer Vision – European Conference on Computer Vision (ECCV) 2014,* 8693:740–755. Lecture Notes in Computer Science. Zurich, Switzerland: Springer. (Cited on pages 41, 43, 54, 71, 78, 125, 130).

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. *DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* 6691–6706. Online: Association for Computational Linguistics. (Cited on pages 26, 33, 117, 126, 129).

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. *Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning*. In *Twelfth International Conference on Learning Representations (ICLR 2024),* 1–45. Vienna, Austria. (Cited on pages 68 sqq.).

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. "A survey on hallucination in large vision-language models," arXiv: 2402.00253. (Cited on pages 2, 4, 32, 69, 74, 76, 124, 129).

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. *Improved Baselines with Visual Instruction Tuning*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 26296–26306. Seattle, Washington, USA: IEEE. (Cited on pages 67, 74).

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. *Visual instruction tuning*. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023),* 1–25. New Orleans, Louisiana, USA. (Cited on pages 2, 17, 22, 67 sq., 130).

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. *P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks*. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Short Papers,* 61–68. Dublin, Ireland: Association for Computational Linguistics. (Cited on pages 22, 31).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv: 1907.11692. (Cited on pages 52, 55).

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. *ParaDetox: Detoxification with parallel data*. In *Proceedings of the 60th Annual Meeting of the Association for Computational,* 6804–6818. Dublin, Ireland: Association for Computational Linguistics. (Cited on pages 5, 33, 87, 124, 130).

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks.* In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019),* 13–23. Vancouver, Canada. (Cited on pages 18, 40, 52, 55 sq., 61).

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. *Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks.* In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 16235–16250. Toronto, Canada: Association for Computational Linguistics. (Cited on pages 86 sq., 130).

Steven G. Luke and Kiel Christianson. 2018. "The Provo Corpus: A large eye-tracking corpus with predictability norms." *Behavior research methods* 50:826–833. (Cited on pages 29, 110, 112, 130).

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. "An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning," arXiv: 2308.08747. (Cited on pages 32, 115).

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. *OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge.* In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 3195–3204. Long Beach, California, USA: IEEE. (Cited on page 78).

Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. 2021. "Fine-Grained Visual Textual Alignment for Cross-Modal Retrieval Using Transformer Encoders." *ACM Transactions on Multimedia Computing, Communications, and Applications* 17 (4): 128:1–128:23. (Cited on page 40).

Meta AI. 2024. *Introducing Meta Llama 3: The most capable openly available LLM to date.* (Cited on page 86).

Emiel Miedema, Sabine Waschull, and Christos Emmanouilidis. 2026. "Towards Trustworthy Artificial Intelligence for Decision-Making: A Lifecycle Perspective on Knowledge- and Data-Driven Artificial Intelligence Systems." *Computers in Industry* 174:104409. (Cited on page 2).

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. *InScript: Narrative Texts Annotated with Script Information.* In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* 3485–3493. Portorož, Slovenia: European Language Resources Association (ELRA). (Cited on page 49).

Byung-Doh Oh and William Schuler. 2023. *Transformer-Based Language Model Surprisal Predicts Human Reading Times Best with About Two Billion Training Tokens.* In *Findings of the Association for Computational Linguistics: EMNLP 2023,* 1915–1921. Singapore: Association for Computational Linguistics. (Cited on page 111).

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. *Im2text: Describing images using 1 million captioned photographs.* In *Advances in Neural Information Processing Systems 24 (NeurIPS 2011),* 1–9. Granada, Spain. (Cited on page 61).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, and Alex Ray. 2022. "Training language models to follow instructions with human feedback." *Advances in Neural Information Processing Systems* 35. (Cited on pages 33, 109, 126).

Gustavo Paetzold and Lucia Specia. 2016a. *Semeval 2016 Task 11: Complex Word Identification*. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* 560–569. San Diego, California, USA: Association for Computational Linguistics. (Cited on page 39).

———. 2016b. *Unsupervised Lexical Simplification for Non-native Speakers*. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 30:3761–3767. 1. Phoenix, Arizona, USA: AAAI Press. (Cited on page 39).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics. (Cited on page 78).

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. *Towards Making the Most of ChatGPT for Machine Translation*. In *Findings of the Association for Computational Linguistics: EMNLP 2023,* edited by Houda Bouamor, Juan Pino, and Kalika Bali, 5622–5633. Singapore: Association for Computational Linguistics. (Cited on page 67).

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. *Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models*. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV),* 2641–2649. Santiago, Chile: IEEE. (Cited on pages 4, 30 sq., 54, 130).

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. "Adapters: A unified library for parameter-efficient and modular transfer learning," arXiv: 2311.11077. (Cited on pages 31 sq., 110 sq., 127).

Tao Qin. 2020. *Dual Learning*. Springer. (Cited on page 53).

Qwen Team. 2026. *Qwen3.5: Accelerating Productivity with Native Multimodal Agents*, February. (Cited on page 20).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack Clark. 2021. *Learning Transferable Visual Models From Natural Language Supervision*. In *Proceedings of the 38th International Conference on Machine Learning,* 139:8748–8763. Online: PMLR. (Cited on pages 2, 18 sq., 30 sq., 41, 52, 55 sq., 61, 69, 124, 127, 129).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI blog,* (cited on pages 16, 26, 32, 112).

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. *Direct preference optimization: Your language model is secretly a reward model*. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023),* 1–27. New Orleans, Louisiana, USA. (Cited on pages 33, 126, 129).

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. 2025. "A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models," arXiv: 2407.02646. (Cited on pages 34, 111, 125).

Jun Rao, Fei Wang, Liang Ding, Shuhan Qi, Yibing Zhan, Weifeng Liu, and Dacheng Tao. 2022. *Where Does the Performance Improvement Come From? – A Reproducibility Concern about Image-Text Retrieval*. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2727–2737. Madrid, Spain: Association for Computing Machinery. (Cited on pages 52, 55 sq., 61).

Keith Rayner. 1998. "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin* 124:372–422. (Cited on pages 28, 34, 110, 112, 126 sq., 129).

Keith Rayner and Alexander Pollatsek. 2016. "Eye movements in reading a tutorial review." *Attention and performance XII,* 327–362. (Cited on page 28).

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015),* 91–99. Montréal, Canada. (Cited on pages 40, 44).

Irina Rets and Jekaterina Rogaten. 2021. "To Simplify or Not? Facilitating English L2 Users' Comprehension and Processing of Open Educational Resources in English using Text Simplification." *Journal of Computer Assisted Learning* 37 (3): 705–717. (Cited on page 37).

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. *Object Hallucination in Image Captioning*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* 4035–4045. Brussels, Belgium: Association for Computational Linguistics. (Cited on pages 3 sq., 32, 67, 78, 124).

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP." *Transactions of the Association for Computational Linguistics* 9:1408–1424. (Cited on page 33).

Florian Schneider. 2021. *Self-Supervised Multi-Modal Text-Image Retrieval Methods to Improve Human Reading*. Master's thesis, University of Hamburg. (Cited on page 45).

Florian Schneider, Özge Alaçam, **Xintong Wang**, and Chris Biemann. 2021. *Towards Multi-Modal Text-Image Retrieval to Improve Human Reading*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop (SRW at NAACL 2021),* 1–8. Online: Association for Computational Linguistics. (Cited on pages 4, 31).

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. "LAION-400M: Open Dataset of CLIP-filtered 400 Million Image-Text Pairs," arXiv: 2111.02114. (Cited on page 41).

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. *Confident Adaptive Language Modeling*. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022),* 1–17. Online. (Cited on pages 112 sq., 119).

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. *A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge*. In *Computer Vision – European Conference on Computer Vision (ECCV) 2022,* 13668:146–162. Lecture Notes in Computer Science. Tel Aviv, Israel: Springer. (Cited on pages 31, 74).

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." *International Journal of Computer Vision* 128, no. 2 (October): 336–359. (Cited on page 50).

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. *Mitigating Hallucinations and Off-target Machine Translation with Source-Contrastive and Language-Contrastive Decoding*. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers),* 21–33. St. Julian's, Malta: Association for Computational Linguistics. (Cited on pages 72, 112, 126).

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. *Semeval-2021 Task 1: Lexical Complexity Prediction.* In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021),* 1–16. Online: Association for Computational Linguistics. (Cited on page 39).

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. *Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,* 2556–2565. Melbourne, Australia: Association for Computational Linguistics. (Cited on page 61).

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. *Self-Attention with Relative Position Representations.* In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers),* 464–468. New Orleans, Louisiana, USA: Association for Computational Linguistics. (Cited on page 14).

Ibne Farabi Shihab, Sanjeda Akter, and Anuj Sharma. 2026. "Calibrated Adaptation: Bayesian Stiefel Manifold Priors for Reliable Parameter-Efficient Fine-Tuning," arXiv: 2602.17809. (Cited on pages 4, 24, 31 sq., 124 sq., 127).

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. *Towards VQA Models That Can Read*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 8317–8326. Long Beach, California, USA: IEEE. (Cited on page 78).

Ping Song, Adegboyega Ojo, and Edward Curry. 2026. "Trustworthy Requirements for Foundation Models: A Comprehensive Survey and Roadmap." *Engineering Applications of Artificial Intelligence* 163:113111. (Cited on pages 2 sq., 124, 126, 128 sq.).

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. *A Mechanistic Interpretation of Arithmetic Reasoning in Language Models using Causal Mediation Analysis*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing,* 7035–7052. Singapore: Association for Computational Linguistics. (Cited on pages 34, 111).

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. *Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention.* In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* 12964–12974. Vienna, Austria: Association for Computational Linguistics. (Cited on pages 5, 27, 32 sq., 124, 126, 129).

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. *Aligning Large Multimodal Models with Factually Augmented RLHF*. In *Findings of the Association for Computational Linguistics: ACL 2024,* 13088–13110. Bangkok, Thailand: Association for Computational Linguistics. (Cited on pages 32, 68).

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. *VL-Adapter: Parameter-Efficient Transfer Learning for Vision-and-Language Tasks*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 5227–5237. New Orleans, Louisiana, USA: IEEE. (Cited on pages 23, 31 sq., 52).

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. "A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models," arXiv: 2401.01313. (Cited on page 69).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. 2023. "LLaMA: Open and Efficient Foundation Language Models," arXiv: 2302.13971. (Cited on pages 52, 55, 67, 112, 124).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, and Shruti Bhosale. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models," arXiv: 2307.09288. (Cited on page 67).

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. *CEFR-based Lexical Simplification Dataset*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* 1–5. Miyazaki, Japan: European Language Resources Association (ELRA). (Cited on page 40).

Sowmya Vajjala and Ivana Lučić. 2018. *OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification*. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications,* 297–304. New Orleans, Louisiana, USA: Association for Computational Linguistics. (Cited on page 49).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention Is All You Need*. In *Advances in Neural Information Processing Systems,* 30:5998–6008. Long Beach, California, USA. (Cited on pages 14 sqq., 40, 127).

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. *CIDEr: Consensus-based Image Description Evaluation*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 4566–4575. Boston, Massachusetts, USA: IEEE. (Cited on page 78).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP,* 353–355. Brussels, Belgium: Association for Computational Linguistics. (Cited on page 116).

Fei Wang, Liang Ding, Jun Rao, Ye Liu, Li Shen, and Changxing Ding. 2024. "Can Linguistic Knowledge Improve Multimodal Alignment in Vision-Language Pretraining?" *ACM Transactions on Multimedia Computing, Communications, and Applications* 20 (12). (Cited on page 68).

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023. *Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 19175–19186. Vancouver, Canada: IEEE. (Cited on pages 18, 52, 55, 69).

Xintong Wang, Xiaoyu Li, Liang Ding, Sanyuan Zhao, and Chris Biemann. 2023. *Using Self-Supervised Dual Constraint Contrastive Learning for Cross-Modal Retrieval*. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023),* 2552–2559. Kraków, Poland: IOS Press. (Cited on pages 10 sq., 32, 69, 111, 125).

Xintong Wang, Xiaoyu Li, Xingshan Li, and Chris Biemann. 2024. *Probing Large Language Models from a Human Behavioral Perspective*. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING 2024,* 1–7. Torino, Italia: ELRA / ICCL. (Cited on pages 11, 34, 111, 126).

Xintong Wang, Yixiao Liu, Jingheng Pan, Liang Ding, Longyue Wang, and Chris Biemann. 2025. *Chinese Toxic Language Mitigation via Sentiment Polarity Consistent Rewrites*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* 35695–35711. Suzhou, China: Association for Computational Linguistics. (Cited on pages 5, 10 sq., 33, 125 sq.).

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. *Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding*. In *Findings of the Association for Computational Linguistics: ACL 2024,* 15840–15853. Bangkok, Thailand: Association for Computational Linguistics. (Cited on pages 10 sq., 33, 112, 126).

Xintong Wang, Jingheng Pan, Liang Ding, Longyue Wang, Longqin Jiang, Xingshan Li, and Chris Biemann. 2025. *CogSteer: Cognition-Inspired Selective Layer Intervention for Efficiently Steering Large Language Models*. In *Findings of the Association for Computational Linguistics (ACL 2025),* 25507–25522. Vienna, Austria. (Cited on pages 10 sq., 27, 34, 87, 125 sq.).

Xintong Wang*, Florian Schneider*, Özge Alaçam, Prateek Chaudhury, and Chris Biemann. 2022. *MOTIF: Contextualized Images for Complex Words to Improve Human Reading*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference,* 2468–2477. Marseille, France: European Language Resources Association. (Cited on pages 4, 9 sq., 31, 54, 125).

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. *Finetuned Language Models Are Zero-Shot Learners*. In *Tenth International Conference on Learning Representations (ICLR 2022).* Online. (Cited on page 109).

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. "Emergent Abilities of Large Language Models." *Transactions on Machine Learning Research,* (cited on page 109).

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, and Atoosa Kasirzadeh. 2021. "Ethical and Social Risks of Harm from Language Models," arXiv: 2112.04359. (Cited on pages 2, 33, 128).

Zhihua Wen, Zhiliang Tian, Zexin Jian, Zhen Huang, Pei Ke, Yifu Gao, Minlie Huang, and Dongsheng Li. 2024. *Perception of Knowledge Boundary for Large Language Models through Semi-open-ended Question Answering*. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024).* Vancouver, Canada. (Cited on page 69).

Anton Wiehe, Florian Schneider, Sebastian Blank, **Xintong Wang**, Hans-Peter Zorn, and Chris Biemann. 2022. *Language over Labels: Contrastive Language Supervision Exceeds Purely Label-Supervised Classification Performance on Chest X-Rays*. In *Proceedings of the 2022 Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop (SRW at AACL-IJCNLP 2022),* 76–83. Online: Association for Computational Linguistics. (Cited on page 69).

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. *ToxiCloakCN: Evaluating Robustness of Offensive Language Detection in Chinese with Cloaking Perturbations*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* 6012–6025. Miami, Florida, USA: Association for Computational Linguistics. (Cited on page 87).

Heping Xie, Richard E. Mayer, Fuxing Wang, and Zongkui Zhou. 2019. "Coordinating Visual and Auditory Cueing in Multimedia Learning." *Journal of Educational Psychology* 111:235–255. (Cited on page 37).

Ming Xu. 2023. *Text2vec: Text to vector toolkit.* https://github.com/shibing624/text2vec. (Cited on page 94).

Rongwu Xu, Zian Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024. *Walking in Others' Shoes: How Perspective-Taking Guides Large Language Models in Reducing Toxicity and Bias*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, 8341–8368. Miami, Florida, USA: Association for Computational Linguistics, November. (Cited on page 94).

Ziyang Xu, Keqin Peng, Liang Ding, Dacheng Tao, and Xiliang Lu. 2024. *Take Care of Your Prompt Bias! Investigating and Mitigating Prompt Bias in Factual Knowledge Extraction*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 15552–15565. Torino, Italia: ELRA / ICCL. (Cited on page 84).

Neemesh Yadav, Sarah Masud, Vikram Goyal, Md Shad Akhtar, and Tanmoy Chakraborty. 2024. *Tox-BART: Leveraging Toxicity Attributes for Explanation Generation of Implicit Hate Speech*. In *Findings of the Association for Computational Linguistics: ACL 2024,* edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar, 13967–13983. Bangkok, Thailand: Association for Computational Linguistics, August. (Cited on page 94).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. "Qwen3 Technical Report," arXiv: 2505.09388. (Cited on pages 86, 124, 129).

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and Haoran Wei. 2024. "Qwen2.5 Technical Report," arXiv: 2412.15115. (Cited on pages 88, 109).

Qingpo Yang, Yakai Chen, Zihui Xu, Yu-ming Shang, Sanchuan Guo, and Xi Zhang. 2025. "SCCD: A Session-based Dataset for Chinese Cyberbullying Detection," arXiv: 2501.15042. (Cited on pages 86 sq., 130).

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. 2023. *Aim: Adapting image models for efficient video action recognition*. In *Eleventh International Conference on Learning Representations (ICLR 2023)*. Kigali, Rwanda. (Cited on page 52).

Kai Yao, Zhenghan Song, Kaixin Wu, Mingjie Zhong, Danzhao Cheng, Zhaorui Tan, Yixin Ji, and Penglei Gao. 2026. "GAST: Gradient-Aligned Sparse Tuning of Large Language Models with Data-Layer Selection," arXiv: 2603.09865. (Cited on pages 4, 24, 31 sq., 127).

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2025. "Knowledge Circuits in Pretrained Transformers," arXiv: 2405.17969. (Cited on page 111).

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. *A Report on the Complex Word Identification Shared Task 2018*. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 66–78. New Orleans, Louisiana, USA: Association for Computational Linguistics. (Cited on page 39).

Huifeng Yin, Yu Zhao, Minghao Wu, Xuanfan Ni, Bo Zeng, Hao Wang, Tianqi Shi, Liangying Shao, Chenyang Lyu, and Longyue Wang. 2025. "Marco-o1 v2: Towards Widening The Distillation Bottleneck for Reasoning Models," arXiv: 2503.01461. (Cited on page 109).

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. "Woodpecker: hallucination correction for multimodal large language models." *Science China Information Sciences* 67, no. 12 (December). (Cited on pages 32, 68 sq., 126).

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. *Ferret: Refer and Ground Anything Anywhere at Any Granularity*. In *Twelfth International Conference on Learning Representations (ICLR 2024)*, 1–30. Vienna, Austria. (Cited on pages 68 sqq.).

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. "CoCa: Contrastive Captioners are Image-Text Foundation Models." *Transactions on Machine Learning Research*, (cited on pages 19, 30, 52, 55).

Tengfei Yu, Liang Ding, Xuebo Liu, Kehai Chen, Meishan Zhang, Dacheng Tao, and Min Zhang. 2023. *PromptST: Abstract Prompt Learning for End-to-End Speech Translation*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10140–10154. Singapore: Association for Computational Linguistics. (Cited on page 114).

Zihao Yue, Liang Zhang, and Qin Jin. 2024. *Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11766–11781. Bangkok, Thailand: Association for Computational Linguistics. (Cited on page 78).

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. *Marco-Bench-MIF: On Multilingual Instruction-Following Capability of Large Language Models*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL),* 24058–24072. Vienna, Austria: Association for Computational Linguistics. (Cited on page 109).

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2024. "HallE-Control: Controlling Object Hallucination in Large Multimodal Models," arXiv: 2310.01779. (Cited on page 68).

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, and Yuansheng Ni. 2024. "A Comprehensive Study of Knowledge Editing for Large Language Models," arXiv: 2401.01286. (Cited on pages 34, 112 sq.).

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2024. *LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention*. In *Twelfth International Conference on Learning Representations (ICLR 2024),* 1–30. Vienna, Austria. (Cited on pages 117, 119).

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. "Explainability for large language models: A survey." *ACM Transactions on Intelligent Systems and Technology* 15:1–38. (Cited on pages 3, 34).

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, and Zican Dong. 2025. "A Survey of Large Language Models," arXiv: 2303.18223. (Cited on page 109).

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. *Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions*. arXiv: 2411.14405. (Cited on page 109).

Ren Zhibo, Wang Huizhen, Zhu Muhua, Wang Yichao, Xiao Tong, and Zhu Jingbo. 2023. *Overcoming Language Priors with Counterfactual Inference for Visual Question Answering*. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics,* 600–610. Harbin, China: Chinese Information Processing Society of China. (Cited on pages 68, 70).

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. "Can ChatGPT Understand Too? A Comparative Study on ChatGPT and Fine-tuned BERT," arXiv: 2302.10198. (Cited on page 67).

———. 2024. *ROSE Doesn't Do That: Boosting the Safety of Instruction-Tuned Large Language Models with Reverse Prompt Contrastive Decoding*. In *Findings of the Association for Computational Linguistics: ACL 2024,* 13721–13736. Bangkok, Thailand: Association for Computational Linguistics. (Cited on page 84).

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. *Learning Deep Features for Discriminative Localization*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2921–2929. Las Vegas, Nevada, USA: IEEE. (Cited on page 50).

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. *Towards identifying social bias in dialog systems: Framework, dataset, and benchmark*. In *Findings of the Association for Computational Linguistics: EMNLP 2022,* 3576–3591. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. (Cited on pages 86 sq., 130).

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. "Learning to Prompt for Vision-Language Models." *International Journal of Computer Vision* 130 (9): 2337–2348. (Cited on pages 53, 55, 59).

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. *Analyzing and Mitigating Object Hallucination in Large Vision-Language Models*. In *Twelfth International Conference on Learning Representations (ICLR 2024),* 1–30. Vienna, Austria. (Cited on page 68).

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. *MiniGPT-4: Enhancing vision-language understanding with advanced large language models*. In *Twelfth International Conference on Learning Representations (ICLR 2024),* 1–15. Vienna, Austria. (Cited on pages 2, 69).

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. "InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models," arXiv: 2504.10479. (Cited on page 20).

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. *Overcoming Language Priors with Self-supervised Learning for Visual Question Answering*. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,* 1083–1089. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization. (Cited on pages 68, 70).