UNIVERSITY OF CALIFORNIA SAN DIEGO

Multimodal Understanding, Alignment and Reasoning in Language Models

A dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Computer Science

by

Junda Wu

Committee in charge:

Professor Julian McAuley, Chair

Professor Prithviraj Ammanabrolu
Professor Biwei Huang
Professor Jingbo Shang

2026

The dissertation of Junda Wu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

EPIGRAPH

Between the idea
And the reality
Between the motion
And the act
Falls the Shadow

— *T. S. Eliot,* The Hollow Men

TABLE OF CONTENTS

Dissertation Approval Page . . . iii
Epigraph . . . iv
Table of Contents . . . v
List of Figures . . . ix
List of Tables . . . xi
Acknowledgements . . . xiii
Vita . . . xv
Abstract of the Dissertation . . . xvii
Chapter 1 Introduction . . . 1
1.1 Representation: Multimodal Understanding . . . 2
1.2 Preference: Multimodal Alignment . . . 3
1.3 Inference: Multimodal Reasoning . . . 4
1.4 Three Thrusts, One Program . . . 5
Chapter 2 Coordinated Multimodal Instruction Tuning (CoMMIT) . . . 6
2.1 Introduction . . . 6
2.2 Related Works . . . 9
2.3 Preliminaries . . . 10
2.4 Measurement of Learning Balance in MLLM Instruction Tuning . . . 11
2.5 Empirical Observations of Learning Dilemmas in MLLM Instruction Tuning . . . 12
2.5.1 The Dilemma of Oscillation . . . 13
2.5.2 The Dilemma of Biased Learning . . . 13
2.6 CoMMIT: Coordinated Multimodal Instruction Tuning . . . 14
2.7 Experiment . . . 16
2.8 Theoretical Analysis . . . 21
2.8.1 Setup and Notations . . . 21
2.8.2 Necessary Assumptions . . . 22
2.8.3 Convergence Analysis . . . 22
2.9 Chapter Summary . . . 23
Chapter 3 Modality-Decoupled Gradient Descent (MDGD) . . . 25

3.1 Introduction . . . 25
3.1.1 Visual Knowledge Forgetting in MLLMs . . . 28
3.1.2 Information Theory in LLMs . . . 28
3.2 Preliminary . . . 29
3.3 MDGD: Modality-Decoupled Gradient Regularization and Descent . . . 30
3.3.1 Modality Decoupling . . . 31
3.3.2 Regularized Gradient Descent . . . 32
3.3.3 Enabling Parameter-efficient Fine-tuning of MDGD via Gradient Masking 33
3.4 Experiments . . . 34
3.4.1 Comparison Results . . . 35
3.4.2 Ablation Study . . . 38
3.4.3 Representation Learning Analysis . . . 38
3.4.4 Sensitivity Study . . . 40
3.5 Chapter Summary . . . 40

Chapter 4 Offline Chain-of-Thought Evaluation and Alignment via Knowledge Graph Exploration (OCEAN) . . . 42
4.1 Introduction . . . 42
4.2 Related Work . . . 44
4.3 Preliminary . . . 46
4.3.1 Problem Formulation: CHAIN-OF-THOUGHT AS AN MDP . . . 46
4.3.2 Verbalized Knowledge Graph Reasoning . . . 47
4.4 **OCEAN**: **O**ffline **C**hain-of-thought **E**valuation and **A**lignme**N**t . . . 48
4.4.1 Offline Evaluation and Optimization . . . 48
4.4.2 Knowledge Graph Preference Modeling . . . 50
4.5 Experiments . . . 51
4.5.1 Implementation Details . . . 51
4.5.2 Multi-hop Question Answering . . . 53
4.5.3 Knowledge-intensive Question Answering . . . 53
4.5.4 Commonsense Reasoning . . . 55
4.6 Analysis . . . 56
4.6.1 In-context Learning & Instruction tuning . . . 56
4.6.2 Evaluation of Generation Quality Post Alignment . . . 57
4.6.3 Case Study . . . 58
4.7 Chapter Summary . . . 59

Chapter 5 In-context Ranking Preference Optimization (IRPO) . . . 60

5.1 Introduction . . . 60
5.2 Related Work . . . 62
5.2.1 Ranking Generation with LLMs . . . 62
5.2.2 Direct Preference Optimization for Ranking . . . 62
5.3 Preliminaries . . . 63
5.3.1 Direct Preference Alignment . . . 63
5.3.2 Discounted Cumulative Gain . . . 64
5.4 IRPO: In-context Ranking Preference Optimization . . . 65
5.4.1 Preference Modeling . . . 65
5.4.2 Policy Optimization Objective . . . 66
5.4.3 Gradient Analysis and Theoretical Insights . . . 67
5.5 Experiments . . . 68
5.5.1 Experimental Setup . . . 68
5.5.2 Conversational Recommendation . . . 69
5.5.3 Generative Retrieval . . . 71
5.5.4 Question-answering as Re-ranking . . . 71
5.6 Analysis . . . 72
5.6.1 On-policy Online Optimization of Iterative IRPO . . . 73
5.6.2 Optimization Analysis of IRPO . . . 74
5.6.3 Ablation Study . . . 75
5.6.4 Comparison with Learning-to-Rank (LiPO) . . . 75
5.7 Chapter Summary . . . 76

Chapter 6 Debiasing Chain-of-Thought via Causal Intervention (DeCoT) . . . 77
6.1 Introduction . . . 77
6.2 Related Work . . . 80
6.3 A Causal View . . . 81
6.3.1 SCM for External Knowledge . . . 82
6.3.2 SCM for Chain-of-thought . . . 82
6.4 Preliminaries . . . 83
6.4.1 Task Formulation . . . 83
6.5 DeCoT: Debiasing Chain-of-thought . . . 84
6.5.1 SCM for DeCoT . . . 84
6.5.2 External Knowledge as an Instrumental Variable . . . 85
6.5.3 Average Causal Effect Guided Chain-of-thought Sampling . . . 86
6.6 Experiments . . . 87

6.6.1 Dataset and Evaluation . . . 87
6.6.2 Baseline and Backbone Model . . . 88
6.6.3 Main Results . . . 89
6.6.4 Improving ReAct by DeCoT . . . 89
6.6.5 Impact of the Selected Entities . . . 90
6.6.6 Impact of the Alternative Entities . . . 91
6.6.7 Case Study . . . 91
6.7 Chapter Summary . . . 93
Chapter 7 Chain-of-Thought Reasoning via Latent State Transition (CTRLS) . . . 95
7.1 Introduction . . . 95
7.2 Related Work . . . 97
7.3 Preliminaries . . . 98
7.3.1 Chain-of-thought Reasoning . . . 98
7.3.2 Distributional Reinforcement Learning . . . 99
7.3.3 Formulation: Transition-aware Chain-of-Thought as an MDP . . . 100
7.4 CTRLS: Chain-of-Thought Reasoning via Latent State Transition . . . 101
7.4.1 Mathematical Assumptions . . . 102
7.4.2 Latent State Encoding . . . 102
7.4.3 State-aware Chain-of-thought Alignment . . . 103
7.4.4 On-Policy Chain-of-thought Reinforcement Learning . . . 106
7.5 Experiments . . . 108
7.5.1 Experimental Setup . . . 108
7.5.2 State Transition-aware Exploration . . . 108
7.5.3 Impact of Exploration Configurations . . . 109
7.5.4 Case Study . . . 110
7.6 Chapter Summary . . . 111
Chapter 8 Conclusion and Future Work . . . 113
8.1 Summary of Contributions . . . 113
8.2 Future Directions . . . 115
References . . . 117

LIST OF FIGURES

Figure 2.1. Illustration of (a) the oscillation problem and (b) the biased learning problem, caused by imbalanced multimodal learning. The optimization trajectories are shown in solid bold lines and the multimodal gradients at the current step t are in solid thin lines. 7

Figure 2.2. Learning curves of the variables H_t^S, H_t^X, and κ_t to measure learning balance in multimodal instruction tuning. 13

Figure 2.3. The learning curves of normalized learning gradient $\|G_t^S\|/\|\theta_t^S\|$ and $\|G_t^X\|/\|\theta_t^X\|$ for the feature encoder and language model respectively, as well as the cross-entropy training losses. 14

Figure 2.4. Learning curves of the multimodal learning balance coefficient κ_t for multiple methods. In addition to the learning curve, we also report the standard deviation of κ_t of each method. 17

Figure 2.5. Instruction-tuning learning curves of BLIP-2 on three vision-based downstream tasks. 18

Figure 2.6. Instruction-tuning learning curves of SALMONN on three audio-based downstream tasks. 18

Figure 3.1. The top-10 image tokens with the highest effective ranks on OKVQA and POPE encoded by LLaVA, and PathVQA and POPE encoded by MiniCPM. 25

Figure 3.2. T-SNE plots of the distribution of extracted visual $\pi(X^v)$ and multimodal z^{vl} representations from pre-trained MiniCPM, and models with direct fine-tuning and MDGD on PathVQA and TextCaps. 37

Figure 3.3. T-SNE plots of the distribution of extracted visual $\pi(X^v)$ and multimodal z^{vl} representations from pre-trained LLaVA-1.5, and models with direct fine-tuning and MDGD on OKVQA and Flickr30K. 37

Figure 3.4. Illustration of (a) the learning process of three methods based on task loss $\mathcal{L}_{vl}(\phi, \theta)$, (b) the average regularized cosine similarity $\frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\|\|\bar{g}_\theta\|}$ in Eq.(3.10) for gradient masking at varying ratios, and (c) the visual representation loss $\mathcal{L}_v(\phi, \theta)$ in Eq.(3.5) for gradient masking at varying ratios α. 37

Figure 3.5. The effective rank comparison on individual downstream fine-tuning datasets. 40

Figure 4.1. Sampling distributions of (a) trajectories in the knowledge graph that are verbalized as multi-step QA tasks and successfully answered by the LLM itself, (b) relations, and (c) entities in the knowledge graphs. 50

Figure 4.2. Comparison results of base LLMs and OCEAN on three evaluation metrics, Self-BLEU, Distinct-2, and AlignScore. Lower Self-BLEU scores and higher Distinct-2 scores indicate better diversity of the generated text, while higher AlignScore indicates better faithfulness in the generated answers. . . 57

Figure 4.3. Sample comparison between the base model and OCEAN on Llama-3 and Gemma-2. Our method enables a more precise and concise Chain of thought. 58

Figure 5.1. Comparison of REINFORCE and IRPO on ARC and CommonsenseQA. . . 73

Figure 5.2. IRPO's performance across six benchmarks using Llama3, Gemma2, and Phi3 74

Figure 6.1. LLMs' internal knowledge bias can trigger the usage of irrelevant information in prompts, generate incoherent reasoning chains, and impair the model's logical reasoning ability. 77

Figure 6.2. Structural causal graphs for (a) injection of external knowledge, *i.e.*, context Lewis et al. (2020), (b) chain-of-thought prompting and (c) our approach using external knowledge as an instrumental variable (detailed in Section 6.5.2). 81

Figure 6.3. The illustration of our proposed debiasing chain-of-thought method DeCoT (detailed in Section 6.5). 82

Figure 6.4. Sensitivity studies on the impact of the number of the selected entities and alternative entities. We also include the $T = 0$ and $P = 0$ data points indicating the performance of regular CoT prompting methods. The experiments are conducted on the GPT-3.5 backbone model on the MuSiQue dataset. 91

Figure 7.1. Illustration of the difference between conventional CoT prompting and CTRLS. 96

Figure 7.2. An overview of the proposed two-phase alignment and fine-tuning scheme. 102

Figure 7.3. **Qualitative comparison.** CTRLS correctly verifies candidate solutions and filters out invalid cases, while the baseline fails to check whether the resulting value is truly prime. 110

LIST OF TABLES

Table 2.1. Instruction tuning results for four MLLMs: BLIP-2, SALMONN, InternVL2-8B, and LLaVA-1.5-7B. 20

Table 2.2. Comparison on A-OKVQA, TextVQA, and IconVQA with BLIP-2 backbone model. Baselines are Constant LR that direct fine tune the backbone model with various learning rate. 20

Table 3.1. Performance on various pre-trained tasks of LLaVA-1.5 models fine-tuned on Flickr30K and OKVQA. We report the best performance for each task in a **bold font** while the second best performance underlined. 34

Table 3.2. Performance on various pre-trained tasks of MiniCPM-V2.5 models fine-tuned on PathVQA and TextCaps. We report the best performance for each task in a **bold font** while the second best performance underlined. 36

Table 4.1. Comparison results of `OCEAN`, base LLMs (`Base`), and supervised fine-tuning (`SFT`), on three **Multi-hop Question-answering** tasks. 54

Table 4.2. Comparison results of `OCEAN`, base LLMs (`Base`), and supervised fine-tuning (`SFT`), on three **Knowledge-intensive Question-answering** tasks. 55

Table 4.3. Comparison results of `OCEAN`, base LLMs (`Base`), and supervised fine-tuning (`SFT`), on five **Commonsense Reasoning** tasks. We report the Exact Match (EM) metric on these tasks and the average performance. We highlight the best method in **bold font** for each task and LLM. 56

Table 4.4. Performance Comparison of In-Context Learning and Instruction Tuning. All datasets consist of classification tasks or true/false questions, so accuracy is used to evaluate the performance. The performance of the base model and our proposed model is largely comparable. 57

Table 5.1. Performance on Redial and Inspired of Conversational Recommendation, evaluated using NDCG and Recall for the top 1, 5, and 10 predictions out of 20 candidate items. 69

Table 5.2. Performance on HotpotQA and MuSiQue for Generative Retrieval, evaluated using NDCG and Recall for the top 1, 3, and 5 predictions out of 10 candidate contexts. 70

Table 5.3. Performance on the ARC and CommonsenseQA QA datasets, evaluated using NDCG and Recall for the top 1, 3, and 5 predictions out of 10 answer choices. 72

Table 6.1. The comparison results of DeCoT based on different backbone LLMs on four knowledge-intensive tasks. The best results for each backbone model and each dataset are highlighted in a **bold font**. 87

Table 6.2. Performance of DeCoT and baselines using retrieved knowledge and reasoning paths from ReAct. Following Yang et al. (2018); Thorne et al. (2018); Yao et al. (2023c), we adopt the EM metric for HotpotQA and the Acc metric for FEVER in the evaluations. 90

Table 6.3. Examples of failure CoTs generated from regular CoT prompting, CoT prompting without context, and knowledge post-edited CoT prompting, as well as DeCoT sampled successful CoTs, from four datasets with the GPT3.5 model. 92

Table 7.1. Comparison of exploration accuracy and success rate of CTRLS and the base models. 108

Table 7.2. RL performance under entropy regularization (left block) and ϵ-greedy exploration (right block). 109

Table 7.3. GPT-4 rubric scores (0–10) on GSM8K with Qwen. Averaged over 100 samples per setting. We evaluate on four different aspects: (**S**)elf-reflection, (**A**)lgebra, (**C**)oherence, (**H**)allucination. 110

ACKNOWLEDGEMENTS

I am grateful to my advisor, Professor Julian McAuley, whose guidance and mentorship have shaped my growth as a researcher. His confidence in allowing me to pursue the questions I found most meaningful gave me the freedom to develop my own research voice, while his insight, patience, and encouragement helped me navigate the many challenges of doctoral study.

I am also sincerely appreciative of my dissertation committee, Professor Jingbo Shang, Professor Prithviraj Ammanabrolu, and Professor Biwei Huang, for their time, careful reading, and thoughtful advice. Their constructive feedback and diverse perspectives helped refine the ideas in this dissertation and strengthened the work in important ways.

This dissertation also reflects the influence of many collaborators, mentors, and colleagues with whom I have been fortunate to work. I am deeply grateful for several years of fruitful collaboration and shared intellectual growth with Xintong Li, whose kindness, trust, and companionship have made this journey especially meaningful. Thanks to my UCSD co-authors, Rohan Surana, Sheldon Yu, Gagan Mundada, Zihan Huang, Yuxin Xiong, Xunyi Jiang, Yash Vishe, Zhouhang Xie, Yu Xia, Yiran Shen, Zachary Novack, Herman Dong, and Yu Wang, who would believe me without a doubt and take on any projects of my wild ideas. Thanks to my Adobe Research colleagues, Tong Yu, Ritwik Sinha, Ryan Rossi, Sungchul Kim, Ruiyi Zhang, Xiang Chen, Rui Wang, and Haoliang Wang, who guided and supported my research. Also, it's a pleasure working with all my other collaborators for their advice and help. Finally, I would like to express my heartfelt appreciation to the people whose love, friendship, and encouragement sustained me throughout this journey.

The fragments of my insights, memories, and emotions remain suspended in the last, lingering light of San Diego's sunset. Beyond flaming clouds, there are my struggling, ecstasy and silence—my eternity, and a day.

Chapter 2, in full, is a reprint of the material as it appears in "CoMMIT: Coordinated Instruction Tuning for Multimodal Large Language Models" by Xintong Li, Junda Wu, Tong Yu, Rui Wang, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Julian McAuley, and Jingbo Shang, pub-

lished in *Proceedings of EMNLP*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "Mitigating Visual Knowledge Forgetting in MLLM Instruction-tuning via Modality-decoupled Gradient Descent" by Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan Rossi, Lina Yao, Jingbo Shang, and Julian McAuley, published in *Findings of EMNLP*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in "Offline Chain-of-Thought Evaluation and Alignment via Knowledge Graph Exploration" by Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, Jingbo Shang, and Julian McAuley, published in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of the material as it appears in "In-context Ranking Preference Optimization (IRPO)" by Junda Wu, Rohan Surana, Zhouhang Xie, Yiran Shen, Yu Xia, Tong Yu, Ryan Rossi, Prithviraj Ammanabrolu, and Julian McAuley, published in *Conference on Language Modeling (COLM)*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of the material as it appears in "DeCoT: Debiasing Chain-of-Thought for Knowledge-Intensive Tasks in Large Language Models via Causal Intervention" by Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan A. Rossi, Sungchul Kim, Anup Rao, and Julian McAuley, published in *Proceedings of ACL*, 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 7, in full, is a reprint of the material as it appears in "CTRLS: Chain-of-Thought Reasoning via Latent State Transition" by Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley, AISTATS, 2026. The dissertation author was the primary investigator and author of this paper.

VITA

2023 Master of Science in Computer Engineering, New York University

2026 Doctor of Philosophy in Computer Science, University of California San Diego

PUBLICATIONS

Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. "CTRLS: Chain-of-Thought Reasoning via Latent State Transition." *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A. Rossi, Julian McAuley, Yunyao Li, and Ritwik Sinha. "Doc-ReAct: Multi-Page Heterogeneous Document Question-Answering." *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Xintong Li, Junda Wu, Tong Yu, Rui Wang, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Julian McAuley, and Jingbo Shang. "CoMMIT: Coordinated Instruction Tuning for Multimodal Large Language Models." *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan Rossi, Lina Yao, Jingbo Shang, and Julian McAuley. "Mitigating Visual Knowledge Forgetting in MLLM Instruction-tuning via Modality-decoupled Gradient Descent." *Findings of EMNLP*, 2025.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, Jingbo Shang, and Julian McAuley. "Offline Chain-of-Thought Evaluation and Alignment via Knowledge Graph Exploration." *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Junda Wu, Rohan Surana, Zhouhang Xie, Yiran Shen, Yu Xia, Tong Yu, Ryan Rossi, Prithviraj Ammanabrolu, and Julian McAuley. "In-context Ranking Preference Optimization." *Conference on Language Modeling (COLM)*, 2025.

Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. "CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation." *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. "InstructGraph:

Boosting Large Language Models via Graph-centric Instruction Tuning and Preference Alignment." *Findings of the Association for Computational Linguistics (ACL Findings)*, 2024.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan A. Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. "DeCoT: Debiasing Chain-of-Thought for Knowledge-Intensive Tasks in Large Language Models via Causal Intervention." *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

ABSTRACT OF THE DISSERTATION

Multimodal Understanding, Alignment and Reasoning in Language Models

by

Junda Wu

Doctor of Philosophy in Computer Science

University of California San Diego, 2026

Professor Julian McAuley, Chair

Multimodal large language models (MLLMs) have demonstrated impressive capabilities across vision-language tasks, yet fundamental challenges remain in understanding their internal multimodal processing, aligning their behavior with human preferences, and enabling faithful reasoning. This dissertation addresses these challenges through three lines of work spanning multimodal understanding, alignment, and reasoning.

We begin with multimodal understanding, where we develop an information-theoretic framework that traces how visual information is encoded, compressed, and shaped by textual instructions across MLLM layers. Building on these insights, we introduce CoMMIT, a coordinated instruction tuning approach that resolves the learning imbalance between visual encoders and the backbone

LLM, and MDGD, which preserves visual knowledge during fine-tuning by decoupling modality-specific gradient updates. We then turn to multimodal alignment, proposing OCEAN, which models chain-of-thought reasoning as a Markov decision process and leverages knowledge graph exploration for offline evaluation and alignment, and IRPO, which extends preference optimization to in-context ranking lists. Finally, we advance multimodal reasoning through DeCoT, which applies causal intervention to debias chain-of-thought reasoning, and CTRLS, which formulates reasoning as latent state transitions with distributional reinforcement learning for structured exploration of diverse reasoning trajectories.

Chapter 1

Introduction

Multimodal large language models (MLLMs) have emerged as a central paradigm in modern artificial intelligence, combining the linguistic competence of large language models with the perceptual capacity to process visual signals, knowledge graphs, ranked feedback, and other non-textual information. Their rapid progress on benchmarks across vision–language understanding, agentic decision-making, and complex reasoning has been impressive; the open research questions, however, do not reduce to scaling alone. They cluster into three thrusts that correspond to the three traditional aspects of an intelligent system: how it *represents* the world, how it *aligns* with the user, and how it *reasons* from one to the other.

This dissertation pursues those three thrusts in turn: *representation* of multimodal input in the language model (Section 1.1), *preference* learning under realistic feedback signals (Section 1.2), and *inference* as structured search over reasoning trajectories (Section 1.3). Each thrust corresponds to one classical face of the multimodal language-model problem—*multimodal understanding*, *multimodal alignment*, and *multimodal reasoning*—and each is addressed by two methodologically complementary contributions. Section 1.4 synthesizes the three thrusts and lays out the remainder of the dissertation.

Each subsequent chapter is a self-contained adaptation of a published or accepted paper, with the original contribution preserved verbatim except for unified notation and stripped reviewer annotations. The closing Conclusion synthesizes findings across the three thrusts, identifies open problems, and sketches several near-term research directions.

1.1 Representation: Multimodal Understanding

The first thrust concerns *representation*: how a language model encodes non-textual signal and whether that representation is preserved as the model is adapted to new tasks. Multimodal large language models are designed so that a feature encoder converts visual or structured input into embedding tokens, which are then inserted into the language prompt and processed by the backbone language model. Effective multimodal behavior depends jointly on the encoder—which must extract task-relevant features from the non-textual modality—and on the backbone—which must adapt its pre-trained capabilities to interpret these features. A growing body of work documents that vanilla instruction tuning silently degrades this multimodal representation: textual loss drives the gradient, while visual information is compressed one way through the encoder and receives no direct supervisory signal. The result is a model that scores well on text-heavy benchmarks while the visual capability the architecture was built for is quietly eroded.

Chapter 2 introduces **CoMMIT**, a coordinated instruction-tuning recipe that operates on the *training-time* side of the problem. A multimodal balance coefficient κ measures the relative learning of the visual encoder and the backbone language model at each step. When κ drifts away from a target regime—one half of the model taking too much gradient—the schedule is adjusted so the slower side catches up. The mechanism is model-agnostic: it depends only on per-component learning quantities, not on architectural specifics, and so it transfers across MLLM backbones such as LLaVA and MiniCPM. CoMMIT also comes with a theoretical convergence analysis that explains why coordinated learning alleviates both the oscillation problem and the biased learning problem.

Chapter 3 introduces **MDGD**, which operates on the *representation-geometry* side rather than the schedule side. Gradient updates are decoupled by modality: language-side updates flow as usual, while visual-side updates pass through a projection that preserves the visual subspace defined by the pre-trained encoder. The resulting effective rank of the visual representation is tracked as a diagnostic and remains stable through fine-tuning instead of collapsing as it does under

vanilla updates. Crucially, MDGD does not require knowing the per-modality balance: it acts on the gradient geometry directly. Where CoMMIT corrects the imbalance once detected, MDGD designs an update that resists the imbalance in the first place. Taken together, the two chapters answer the representation thrust from training-time and representation-side directions.

1.2 Preference: Multimodal Alignment

The second thrust concerns *preference*: how a language model is aligned to user intent given that direct human feedback at scale is impractical. The frontier answer has been Reinforcement Learning from Human Feedback (RLHF), trained on pairwise preferences elicited from human annotators. Pairwise preference, however, is among the sparsest signals one can imagine: it collapses an entire space of possible alternatives, rationales, and degrees of preferability into a single bit. The signal granularity that reaches the model is dramatically narrower than the preference granularity its users actually carry. Existing approaches have treated the optimization objective as the focus of innovation—DPO replaces PPO, simulated preference data replaces human labels, online iterations replace offline ones—but the pattern is consistent: the optimizer changes; the signal does not. The two chapters in this thrust shift the focus from objective to *signal* itself, enriching where the preference signal comes from and how it is shaped.

Chapter 4 introduces **OCEAN**, which addresses the signal *source*. Chain-of-thought reasoning is reformulated as a Markov decision process over knowledge-graph states, with rewards derived from knowledge-graph exploration rather than collected from humans. The dissertation proves an unbiased KG-IPS estimator for off-policy alignment, with an accompanying variance bound that characterizes when the structured world signal carries enough information to align reasoning behavior. The contribution is not a better optimizer but a better reward—one drawn from a structured external source that humans rarely produce in sufficient volume.

Chapter 5 introduces **IRPO**, which addresses the signal *shape*. Pairwise preferences are extended to in-context ranked lists with positional weighting: instead of asking which of A and B is preferred, the framework learns from rankings of k items where position carries information.

Gradient analysis shows where ranking surfaces preference structure that pairwise reduction discards, and empirical results on question answering, conversational recommendation, and generative retrieval-augmented generation show consistent gains over pairwise DPO at matched compute. Where OCEAN enriches what the signal is *about*, IRPO enriches what the signal *is*.

1.3 Inference: Multimodal Reasoning

The third thrust concerns *inference*: how a language model reasons through multi-step problems and whether the surface reasoning trace faithfully represents the underlying computation. Chain-of-thought (CoT) reasoning has become the standard mechanism by which language models externalize their intermediate steps, yet a growing body of work shows that the told-trace and the actual computation can diverge: the model gestures at one path while producing answers on another. Existing remedies have largely accepted CoT as a left-to-right autoregression and intervened around it—more demonstrations, more refinement, more verification passes—but they leave intact the underlying assumption that reasoning is a single privileged trajectory. The two chapters in this thrust depart from that assumption in opposite directions: the first treats spurious paths as a causal artifact to be removed; the second treats reasoning itself as a structured latent space to be explored.

Chapter 6 introduces **DeCoT**, a causal-intervention procedure for knowledge-intensive reasoning. A structural causal model writes down where the model's parametric knowledge and externally retrieved evidence interact, and front-door adjustment over retrieved evidence breaks the confounding path that produces spurious-but-fluent reasoning. The intervention is light: it does not retrain the model and does not require labeled examples of correct reasoning. DeCoT exposes which of the retrieved facts the answer actually relies on, and which spurious detours have been pruned, yielding consistent gains across knowledge-intensive multi-hop QA benchmarks.

Chapter 7 introduces **CTRLS**, which reformulates CoT as a sequence of transitions in a latent state space. An evidence lower bound objective trains a latent transition model jointly with the language model, and distributional reinforcement learning over the latent MDP explores multiple plausible reasoning trajectories rather than collapsing onto one. Entropy regularization controls the

exploration–exploitation trade-off; the resulting reasoning is more diverse, more controllable, and consistently more accurate on mathematical reasoning benchmarks across Llama and Qwen backbones. Where DeCoT *removes* what should not be present in a reasoning path, CTRLS *structures* the space of paths that should be searched.

1.4 Three Thrusts, One Program

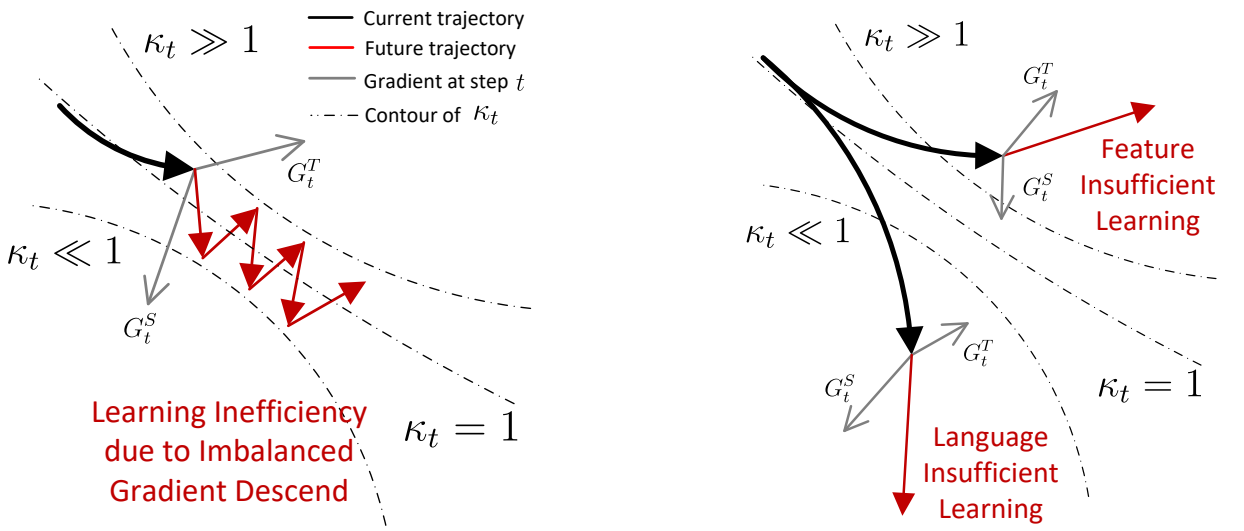
Across the three thrusts, three threads recur and provide the thesis to which the Conclusion returns. Modality-aware tuning is not a refinement but a prerequisite for any language model that ingests non-textual signal. Alignment can move offline without losing signal granularity, provided the source and shape of feedback are designed deliberately. Reasoning is a structured search problem, not a left-to-right autoregression—and modeling its structure changes both diagnosis and improvement. The Conclusion (Chapter 8) synthesizes these three threads, names the open problems that fall out of the present argument, and sketches the next steps—chief among them, turning the diagnostic information-theoretic lens into a constructive one, extending offline alignment beyond curated knowledge graphs, and connecting latent-state reasoning to inference-time compute scaling.

Chapter 2

Coordinated Multimodal Instruction Tuning (CoMMIT)

2.1 Introduction

Multimodal instruction tuning aligns pre-trained multimodal large language models (MLLMs) with specific downstream tasks by fine-tuning MLLMs to follow arbitrary instructions (Dai et al., 2024; Zhang et al., 2023a; Zhao et al., 2024a; Lu et al., 2023; Han et al., 2023; Wu et al., 2024c; Wang et al., 2024c; Wu et al., 2025d,a). Leading pre-trained Multimodal Large Language Models (MLLMs) typically share similar architectures (Li et al., 2023b; Liu et al., 2023a; Tang et al., 2023a; Chu et al., 2023a). Specifically, the non-text data (image, audio, etc) is first encoded by a feature encoder into embedding tokens. Then, these encoded embeddings are inserted into language prompts, creating multimodal sequence inputs for the LLMs. Effective multimodal understanding and reasoning in MLLMs depend on the model's ability to learn aligned multimodal features using its feature encoder (*e.g.*, Li et al. (2023b)), and on leveraging the pre-trained capabilities of its backbone LLM (*e.g.*, Touvron et al. (2023); Chiang et al. (2023)) to interpret these multimodal inputs. This generally involves a two-prolonged learning process: (1) **LLM Adaptation.** Encoded non-text features (*e.g.*, visual and auditory) in downstream tasks may not be perfectly aligned with pre-trained text features, thus requiring the backbone LLM to adapt its pre-trained parameters to



(a) (b)

Figure 2.1: Illustration of (a) the oscillation problem and (b) the biased learning problem, caused by imbalanced multimodal learning. The optimization trajectories are shown in solid bold lines and the multimodal gradients at the current step t are in solid thin lines.

recognize these new, non-text modality tokens. (2) **Feature Encoder Adaptation.** While LLMs possess strong reasoning ability from their pre-trained, it requires the feature encoders to be fine-tuned to extract task-specific information for evidence of reasoning. Cooperative balancing of these two learning stages is crucial for effective instruction tuning of MLLMs. When the learning is biased on the LLMs with (1), the insufficiently learned feature encoder can lead to information loss (Bai et al., 2024; Tong et al., 2024; Wu et al., 2025e), hindering the LLM's ability to reason effectively due to a lack of adequate evidence from non-text modalities. Conversely, if the learning is biased on the feature encoder with (2), the LLMs can be insufficiently adapted and struggle to interpret non-text modalities. As a result, it will cause the hallucination problem (Bai et al., 2024; Rawte et al., 2024; Wu et al., 2024b,f) due to the strong language prior inherent in the backbone LLMs. Therefore, it is essential to balance the learning between the feature encoder and backbone LLM, so that the learning is not overly biased on either of the two modules.

In this paper, we first propose a multimodal balance coefficient that quantifies the learning balance between the feature encoder and the backbone LLM in MLLM instruction tuning. Based on theoretical analysis and empirical observations, we identify two types of learning dilemmas that can be quantitatively measured by our proposed multimodal balance coefficient: *i)* the oscillation

problem and *ii)* the biased learning problem, as illustrated in Figure 2.1. Specifically, Figure 2.1a demonstrates the oscillation problem where the learning is *alternatively* favoring either the feature encoder or the LLM. This oscillation impedes the convergence of optimization and undermines learning efficiency since the learning is hardly progressing in consistent directions. On the other hand, Figure 2.1b shows the biased learning problem where the training *consistently* favors either the LLM or the feature encoder. In such cases, the gradient descent primarily only updates either the LLM or the feature encoder, resulting in insufficient learning of the other module. This diminishes the effectiveness of gradient descent since the under-trained module (LLM or feature encoder) will not be capable of contributing sufficient information to the generation outputs.

To address these challenges, we propose **Co**ordinated **M**ulti**M**odal **I**nstruction **T**uning (**CoMMIT**), which regularizes the training with a coordinated learning rate scheduler (Section 2.6). This scheduler dynamically adjusts the learning rates of the feature encoder and LLM according to the proposed multimodal balance coefficient, ensuring sufficient gradient descent for both the feature encoder and LLM while mitigating the oscillation problem. We also introduce a regularization loss that promotes larger update steps during training, further alleviating gradient diminishing. We theoretically analyze the convergence rate and demonstrate that we can achieve accelerated convergence when optimizing with **CoMMIT** (Section 2.8). We summarize our main contributions as follows:

- We introduce a theoretical framework to uncover the pitfalls of the learning imbalance problem in MLLM instruction tuning, which can cause MLLM insufficient learning and the oscillation problem.
- Based on the theoretical analysis and empirical observation, we propose **CoMMIT** to balance multimodal learning progress by dynamically coordinating learning rates on the feature encoder and LLM. **CoMMIT** also enforces a gradient regularization that encourages larger step sizes and improves training efficiency.
- Applying **CoMMIT** introduces a novel term in the convergence rate analysis. Theoretical analysis proves that this term is always greater than one, leading to faster convergence. We

also demonstrate that the theorem can be generalized across various optimizers.

- Empirical results on downstream tasks in vision and audio modalities with various LLM backbones show the efficiency and effectiveness of the methods. We demonstrate that **CoMMIT** can better coordinate multimodal learning progress with balanced training between the feature encoder and the LLM.

2.2 Related Works

MLLMs have become a new paradigm to empower multimodal learning with advanced language reasoning capabilities, such as with vision (Li et al., 2023b; Liu et al., 2023a; Wang et al., 2024d; Maaz et al., 2023; Zhang et al., 2023a; Huang et al., 2023; Yan et al., 2024), and audio (Huang et al., 2024a; Tang et al., 2023a; Gardner et al., 2023). Despite good generalizability and zero-shot performance of existing large language models (LLMs), the discrepancy between different modalities can be one of the greatest challenges for LMMs to achieve comparable reasoning performance as LLMs. To bridge the multimodality gap and align with downstream tasks, several works focus on two-fold considerations: feature (modality) alignment and reasoning alignment. The most common approach for feature alignment is to encode the source modality feature to semantic tokens within the LLMs' embedding feature space. By adding the modality-specific tokens (Wang et al., 2024a; Liu et al., 2021a; Zhang et al., 2024a) as soft prompt inputs (Liu et al., 2021b; Xie et al., 2024a; Wu et al., 2023a, 2024e), the backbone LLMs can process these tokens with language tokens as a unified sequence. However, the newly added semantic tokens cannot be understood by LLMs directly for language reasoning, due to the limited text-only pretraining of LLMs. Such misalignment problems will lead to textual hallucination problems, namely *linguistic bias* (Ko et al., 2023; Tang et al., 2023b), in which the language models reason only based on their language prior. Thus, multimodal alignment should be achieved by additional adaptation of the LLM itself with multimodal instruction tuning.

2.3 Preliminaries

Given a pair of non-text input I^S (images, audio, etc) and instruction prompt I^X of nature language, the instruction-tuned MLLM should comprehend the semantics of I^S and generate outputs that comply with the instruction specified in I^X. State-of-the-art MLLM instruction tuning generally adopts similar diagrams of training (Gardner et al., 2023; Li et al., 2023b; Liu et al., 2023a), which involves cooperative training between a feature encoder S and a pretrained LLM X. Specifically, S first encodes the multimedia input I^S into the embedding space of X. Then, the encoded I^S is inserted into the instruction prompt I^X as input that conditions the output generation of X,

$$P_{S,X}(\hat{y}_k|I^S, I^X, y_{j<k}) = X(\hat{y}_k|S(I^S), I^X, y_{j<k}), \tag{2.1}$$

where $\hat{y}_k$ is the kth predicted token and $y_{j<k}$ denotes the first $k-1$ of the expected ground truth tokens in auto-regressive generation, $k = 1, \cdots, K$. The training loss is the cross-entropy defined by the predicted distribution on $\hat{y}_k$ and the kth ground truth token y_k(Liu et al., 2023a; Ouyang et al., 2022),

$$L(Y = \{y_k\}_{k=1}^K \mid I_S, I^T) = -\frac{1}{K}\sum_{k=1}^{K} y_k \log P_{S,X}(\hat{y}_k|I^S, I^X, y_{j<k}), \tag{2.2}$$

The learning objective is to find the optimal X and S by minimizing the loss function. As mentioned in Section 2.1, the learning can either be inefficient by oscillating between the optimization of the two modules or insufficient by biasing on one of S and X. Our goal is to find a balance between the learning of X and S, so to accelerate the convergence while ensuring that both S and X are sufficiently trained.

2.4 Measurement of Learning Balance in MLLM Instruction Tuning

To assess the balance between the updates on X and S, we first measure the significance of each update separately with X and S. Formally, for the t-th step of training, we define $d(P_{X_t,S_t}||P_{X_{t+1},S_t})$ and $d(P_{X_t,S_t}||P_{X_t,S_{t+1}})$ that quantify the significance of updates on X and S, respectively, by measuring the shift in output distributions,

$$d(P_{X_t,S_t}||P_{X_{t+1},S_t}) = \frac{1}{K}\sum_k \mathbb{KL}\left(P_{S_t,X_t}(\hat{y}_k|I^S,I^X)||P_{S_{t+1},X_t}(\hat{y}_k|I^S,I^X)\right), \tag{2.3}$$

$$d(P_{X_t,S_t}||P_{X_t,S_{t+1}}) = \frac{1}{K}\sum_k \mathbb{KL}\left(P_{S_t,X_t}(\hat{y}_k|I^S,I^X)||P_{S_t,X_{t+1}}(\hat{y}_k|I^S,I^X)\right) \tag{2.4}$$

where we use the subscript t to index the trained steps and $\mathbb{KL}(\cdot||\cdot)$ is the KL divergence. Based on (2.3) and (2.4), we define the Multimodal Balance Coefficient that measures the balance between training on X and S.

Definition 2.1 (Multimodal balance coefficient)**.** *For time step t of joint training on X and S, the Multimodal Balance Coefficient κ_t is measured with the respective learning steps on the feature encoder S_t and the LLM X_t.*

$$\kappa_t = \frac{d(P_{X_t,S_t}||P_{X_{t+1},S_t})}{d(P_{X_t,S_t}||P_{X_t,S_{t+1}})}. \tag{2.5}$$

κ_t with a value always near 1 indicates that the learning is balanced between the feature encoder and the LLM. On the contrary, $\kappa_t >> 1$ and $\kappa_t \to 0$ suggests that the learning biased by leaning toward X and S, respectively. κ_t with high variance corresponds to learning that oscillates between optimizing on X or S. To further illustrate this, we derive Theorem 1 which estimated the gradient on X and S during training.

Theorem 2.2. *Let G_t^X and G_t^S be the gradient on X and S at time step t, we can derive the multimodal gradient estimated bounds,*

$$\|G_t^X\| \leq (\kappa_t + 1)H_t^S, \quad \|G_t^S\| \leq (\frac{1}{\kappa_t} + 1)H_t^X, \tag{2.6}$$

where H_t^S and H_t^T represent the individual learning steps of S and X, respectively. These are given by,

$$\begin{aligned} H_t^S &= \left(\|I_t^X\| + \|S_t(I_t^S)\|\right)^{-1} \|logits(P_{X_t,S_{t+1}})\|, \\ H_t^X &= \left\|I_t^S\right\|^{-1} \|logits(P_{X_{t+1},S_t})\|. \end{aligned} \tag{2.7}$$

Within the metric space of probability distribution $(P_{S,X}, d)$, the values of H_t^S and H_t^T are bounded by a finite norm. H_t^S and H_t^T are also lower-bounded, assuming multimodal gradients are not diminishing which we alleviate by proposing a regularization on the gradient in Section 2.6. Therefore, κ_t can account the most for the gradient upper bound in (5.12) so it is suitable to measure the learning imbalance problem.

2.5 Empirical Observations of Learning Dilemmas in MLLM Instruction Tuning

In this section, we illustrate the learning dilemmas described in Section 2.1 by observing the dynamics of κ_t is different experiment settings. The experiment is conducted by fine-tuning the BLIP-2 (Li et al., 2023b) model on TextVQA(Singh et al., 2019), a widely used dataset for visual question question answering (Dai et al., 2024; Yin et al., 2024). To probe the problems of oscillation and learning bias, we consider the following three learning strategies: (1) **Synced LR** is trained by setting the learning rate of both X and S to $1e-4$; (2) **Language LR** $\uparrow$ increases learning rate of X to $1e-3$; (2) **Encoder LR** $\uparrow$ increases the learning rate of S to $1e-3$.

2.5.1 The Dilemma of Oscillation

To quantitatively understand the oscillation problem in MLLM instruction tuning, we show the learning curves (Figure 2.2) of the measurement variables H_t^S, H_t^X, and κ_t proposed in (5.12).

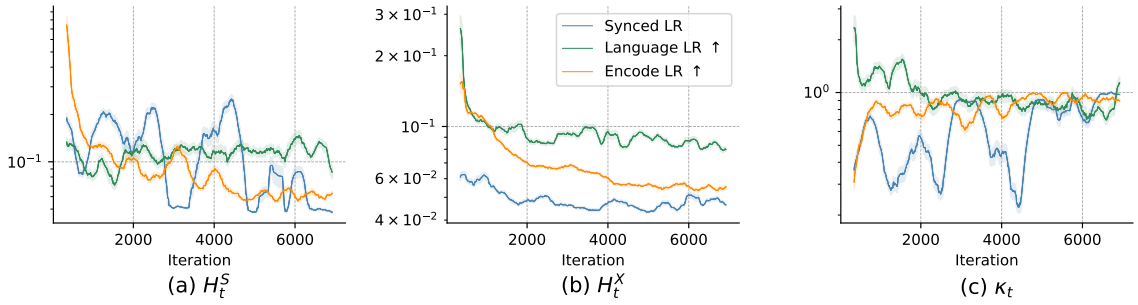


Figure 2.2: Learning curves of the variables H_t^S, H_t^X, and κ_t to measure learning balance in multimodal instruction tuning.

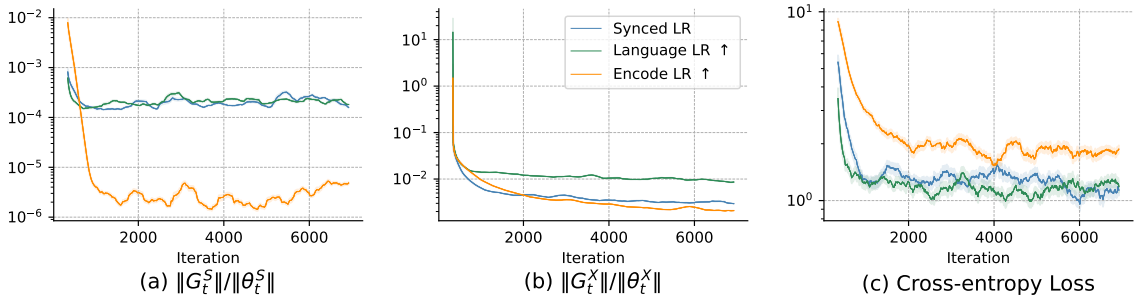
Observation 2.3. *As shown in Figure 2.2(c), the multimodal learning process can suffer from significant oscillation problems with highly variant κ_t in the **Synced LR** setting.*

Specifically, the learning curve of κ_t in the **Synced LR** setting exhibits high variance near the value of 1 , which is a showcase of the oscillation problems that signify training instability. This will cause inefficient training since the learning is alternating between X and S, instead of progressing in consistent directions. We demonstrate such inefficiency in Figures 2.5 and 2.6, showing that its convergence is slower compared to our proposed approach, which balances the learning process with a more stable κ_t. Further, it is interesting to find in Figure 2.2 that the feature encoder S is more unstable than the language model X, by comparing H_t^S in Figure 2.2(a) and H_t^X in Figure 2.2(b). By increasing the learning rate either X (**Encoder LR**) or S (**Language LR**), we observe that the three metrics in Figure 2.2 are stabilized. In the next section, we show that such stabilization is at the expense of biased learning, causing insufficient training on X or S.

2.5.2 The Dilemma of Biased Learning

The **Encoder LR** ↑ and **Language LR** ↑ in Figure 2.2 demonstrated the biased learning problem, where the training is biased on either S or X. Let θ_t^S and θ_t^X be the parameters of S and

X at time step t. We further show three metrics with the same backbone MLLM and training data as in Section 2.5.1: (1) the normalized learning gradient $\|G_t^S\|/\|\theta_t^S\|$ of the feature encoder S in Figure 2.3(a), (2) the normalized learning gradient $\|G_t^X\|/\|\theta_t^X\|$ of the language model X in Figure 2.3(b), and (3) the cross-entropy loss in Figure 2.3(c). These metrics help understand the impact of biased learning on either X or S in MLLM instruction tuning.



(a) $\|G_t^S\|/\|\theta_t^S\|$ (b) $\|G_t^X\|/\|\theta_t^X\|$ (c) Cross-entropy Loss

Figure 2.3: The learning curves of normalized learning gradient $\|G_t^S\|/\|\theta_t^S\|$ and $\|G_t^X\|/\|\theta_t^X\|$ for the feature encoder and language model respectively, as well as the cross-entropy training losses.

Observation 2.4. *In Figure 2.3(c),* ***Encoder LR*** $\uparrow$ *and* ***language LR*** $\uparrow$*, biased on either X or S, can slow the convergence of the MLLM with gradient diminishing and inferior training performance.*

Such diminishing gradient would result in insufficient training on X or S with gradient descent. For example, we can observe in Figure 2.3(a) and Figure 2.3(b) that the **Encoder LR** can simultaneously cause the gradient diminishing in both X and S, with the cross-entropy converging to a higher value in Figure 2.3(c). In such cases, it is necessary to strategically balance the learning between different modules, so the training is not leaning toward either X or S. In Figure 2.4, we show that our proposed approach can reduce such bias while not inducing oscillation, *i.e.*, by learning with less variant κ_t that is valued close to 1.

2.6 CoMMIT: Coordinated Multimodal Instruction Tuning

Based on the observations in Section 2.5, the learning rate on X should be boosted when the training is leaning toward S ($\kappa_t \to 0$), and vice versa. So motivated, we propose a dynamic learning

rate scheduling method to coordinate multimodal learning between X and S, which alleviates the oscillation problems while ensuring the training is not biased on either X or S.

Inspired by damping strategies in optimization (Lucas et al., 2018; Tanaka and Kunin, 2021; Wei et al., 2021), we use the proposed learning balance metric κ_t in (2.5) as the damping parameter that facilitates balanced multimodal learning. Specifically, we track the N_κ moving average of κ_t through the learning process,

$$\tilde{\kappa}_t = \frac{1}{N_\kappa} \sum_{i=1}^{N_\kappa} \kappa_{t-i+1}, \tag{2.8}$$

then dynamically adjust learning rates of X and S in accordance. Let β_t^X and β_t^S be the learning rates on X and S at time step t. We adjust the learning rates by,

$$\beta_t^T = \frac{2\alpha}{\tilde{\kappa}_t + 1}, \ \beta_t^S = \frac{2\alpha}{1/\tilde{\kappa}_t + 1}, \tag{2.9}$$

where α is the base learning rate.

During training, the diminishing H_t^S and H_t^T as observed in Figure 2.3 can cause higher estimation errors in $\tilde{\kappa}_t$. To address these, we propose an auxiliary regularization that encourages large step sizes for both X and S, which mitigates gradient diminishing. Specifically, we want to encourage larger distribution drifts $d(P_{X_t,S_t}||P_{X_{t+1},S_t})$ for X and $d(P_{X_t,S_t}||P_{X_t,S_{t+1}})$ for S, apart from gradient descending on the cross-entropy loss in (2.2). The gradient update for our proposed CoMMIT at time step t is,

$$\theta_{t+1}^X \leftarrow \theta_t^X - \beta_t^X \cdot \nabla_{\theta^X} L(S_t(\tilde{X}_{\theta_t^X})) + \beta_t^X \cdot \nabla_{\theta^X} d(P_{X_t,S_t}||P_{X_{t+1},S_t}), \tag{2.10}$$

$$\theta_{t+1}^S \leftarrow \theta_t^S - \beta_t^S \cdot \nabla_{\theta^S} L(T_t(S_t(X)); \theta_t^S) + \beta_t^S \cdot \nabla_{\theta^S} d(P_{X_t,S_t}||P_{X_t,S_{t+1}}). \tag{2.11}$$

Note that the distribution drifts $d(P_{X_t,S_t}||P_{X_{t+1},S_t})$ and $d(P_{X_t,S_t}||P_{X_t,S_{t+1}})$ does not involve ground truth labels.

Theoretical Convergence Analysis. We further provide a theoretical analysis of our proposed CoMMIT framework in Section 2.8. Specifically, we establish a new convergence bound under sufficient assumptions for the coordinated learning rate and regularization strategy, demonstrating that the proposed method can achieve a provably faster convergence rate compared to standard imbalanced instruction tuning approaches. We further provide the detailed proof of the proposed theorems in the original paper. Theoretical results show that dynamic adjustment of learning rates, guided by the multimodal balance coefficient, consistently accelerates optimization and ensures sufficient learning across modalities. We also discuss how results can be generalized to various gradient-based optimizers.

2.7 Experiment

Experiment Setup We conduct experiments on two non-text modalities, vision and audio, with multiple instruction-tuning downstream tasks: (1) for **Vision**, we evaluate the backbone MLLMs including BLIP-2 (Li et al., 2023b), InternVL2 Chen et al. (2024b), and LLaVA-1.5 Liu et al. (2023a), on three visual question-answering tasks: TextVQA (Singh et al., 2019), IconQA (Lu et al., 2021), and A-OKVQA (Schwenk et al., 2022), which focus on text recognition and reasoning, knowledge-intensive QA, and abstract diagram understanding, respectively; (2) for **Audio**, we leverage the SALMONN (Tang et al., 2023a) model and evaluate one audio question-answering task and two audio captioning tasks: ClothoAQA (Lipping et al., 2022), MACS (Morato and Mesaros, 2021), and SDD (Manco et al., 2023), which focus respectively on crowdsourced audio question-answering, acoustic scene captioning, and text-to-music generation. We include detailed implementation details in the original paper for individual backbone MLLM.

We follow the instruction tuning diagram (Dai et al., 2024; Tang et al., 2023a; Huang et al., 2023), where the parameters of backbone LLMs are finetuned with LoRAs (Hu et al.) and the feature encoders are finetuned directly. We set the learning rate to $1e-4$ for all the feature encoders and backbone LLMs in our baseline methods **Constant LR** (Dai et al., 2024; Tang et al., 2023a), **Feature CD**, **Language CD** (Wright, 2015). For Feature CD, we first update the feature encoder

until its weights stabilize, then update the backbone LLMs. For Language CD, the process is reversed, with the LLMs being trained first. We also use $1e-4$ as the base learning rates for our CoMMIT variants. There are two CoMMIT variants: **CoMMIT** and **CoMMIT-CLR**. **CoMMIT** is out proposed method in this paper, while **CoMMIT-CLR** is an ablation on **CoMMIT**, without the last regularization terms in (2.10) and (2.11).

Mitigating Oscillation and Biased Learning. In Figure 2.4, we inspect the oscillation problem in Section 2.5.1 by showing the value of κ_t with our proposed CoMMIT and CoMMIT-CLR, comparing to the three learning rate scheduling methods described in Section 2.5. We report with the BLIP-2 (Li et al., 2023b) backbone model on the task of TextVQA (Singh et al., 2019). It can be observed that both CoMMIT and CoMMIT-CLR can stabilize multimodal learning with smaller standard deviations of κ_t over time, thus alleviating the oscillation problem in Observation 2.3 (in Section 2.5.1).

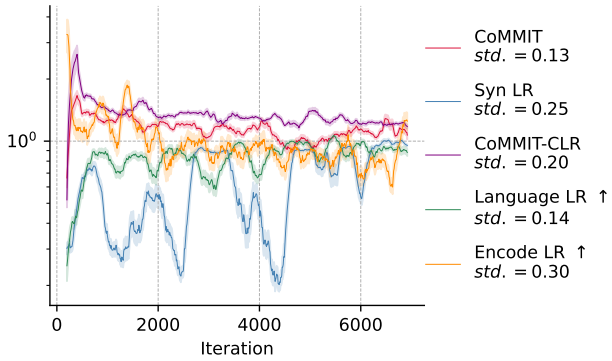
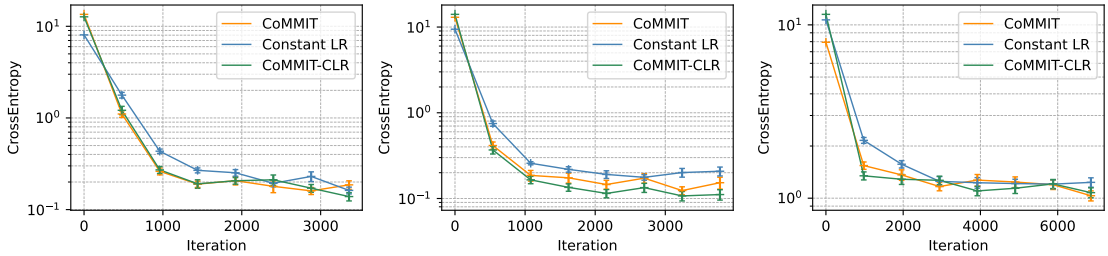


Figure 2.4: Learning curves of the multimodal learning balance coefficient κ_t for multiple methods. In addition to the learning curve, we also report the standard deviation of κ_t of each method.

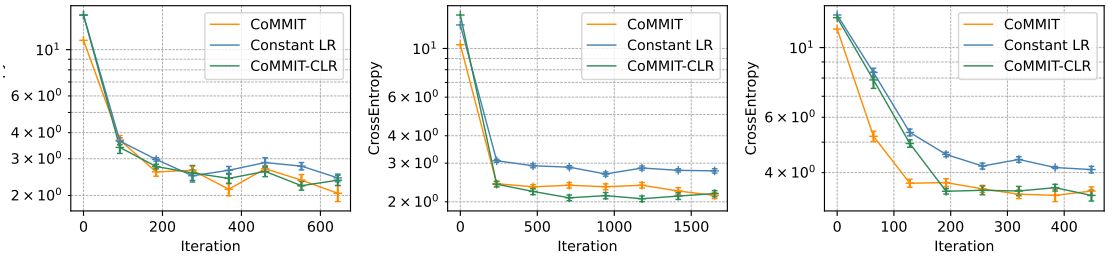
Though Language LR ↑ also yields high stability on κ, such learning rate adjustment method suffers from the problems of biased learning as described in Observation 2.4 (in Section 2.5.2), which potentially causes insufficient training on the feature encoder and worse performance of instruction tuning. Comparing CoMMIT and CoMMIT-CLR, we can observe that CoMMIT achieves less biased learning with the value of κ_t closer to 1 while demonstrating relatively milder learning oscillation with less variant κ_t during training. Further, we find that the proposed loss regularization in Section 2.6 also improves the training balance. Accompanied by the loss regularization

and learning balance coefficient κ_t, CoMMIT more effectively adapts the optimization process for better training between the feature encoder and LLM.



(a) BLIP-2/A-OKVQA (b) BLIP-2/IconQA (c) BLIP-2/TextQA

Figure 2.5: Instruction-tuning learning curves of BLIP-2 on three vision-based downstream tasks.



(a) SALMONN/ClothoAQA (b) SALMONN/MACS (c) SALMONN/SDD

Figure 2.6: Instruction-tuning learning curves of SALMONN on three audio-based downstream tasks.

Note that SynLR with the oscillated value of κ_t is our baseline Constant LR. In Figure 2.5 and 2.6, it is shown that Constant LR generally results in a slower training rate characterized by a flatter descent in loss values during the early stages. This is consistent with Observation 2.3 in Section 2.5.1, suggesting that the oscillation problem can slow the convergence of MLLMs in instruction tuning.

Improved Convergence in MLLM Instruction Tuning. We evaluate the convergence performance of the proposed methods CoMMIT-CLR and CoMMIT in comparison with Constant LR in Figure 2.5 and 2.6. For visual question-answering tasks in Figure 2.5, we observe that CoMMIT-CLR and CoMMIT can accelerate the instruction tuning of BLIP-2 in the early stage. This is especially evident in the case of IconQA, which is out-of-domain for BLIP-2's pretraining. Specifically, IconQA requires regional-level and spatial visual understanding, which differs from the tasks BLIP-2 was pre-trained on (Chen et al., 2023a). In addition, CoMMIT-CLR and CoMMIT achieve

lower training losses than Constant LR, validating CoMMIT's improved training efficiency. This efficiency gain is attributed to CoMMIT's ability to mitigate the oscillation (Section 2.5.1), accelerating convergence.

Similar to the vision-based tasks, we can find in Figure 2.6 that CoMMIT-CLR and CoMMIT can also converge to lower loss values in audio tasks. Specifically, we observe that the CoMMIT-CLR and CoMMIT can achieve better accelerations on the audio captioning tasks of MACS and SDD, compared to training on the audio question-answering task of ClothoAQA. Since audio captioning tasks need more adaptation in MLLMs to generate relatively longer context and align the generation distribution with specific tasks, the coordinated learning rate scheduling method in Section 2.6 can more dynamically adjust the learning rate for less learned components at each model update step. In addition, we show that the proposed loss regularization method adopted in CoMMIT can actively promote the difference in MLLM's generation distribution between optimization steps, which can better benefit tasks, such as audio captioning, that require the model to generate longer contexts. Such improved convergence is associated with our proposed balancing strategy that improves on the oscillation and biased learning problem as discussed.

Improved Downstream Performance across Modalities. In Table 2.1, We evaluate the performance of the proposed methods CoMMIT-CLR and CoMMIT, comparing with three baselines Constant LR, Feature CD, and Language CD. Among the three baselines, we observe that coordinate gradient descend methods have the most improvement compared to the constant learning rate methods that show significant learning tendencies towards a certain modality (*e.g.*, Language CD in SDD, and Feature CD in A-OKVQA and ClothoAQA). However, since such learning balance varies in downstream tasks, coordinate descend methods cannot consistently improve MLLM instruction tuning, while arbitrarily inclining towards only a certain modality can result in inferior model performance (*e.g.*, Feature CD in SDD and Language CD in A-OKVQA).

Different from the fixed learning tendency which needs to be predetermined by coordinate descend methods, the proposed coordinated learning rate scheduling method can dynamically adapt learning rates for multimodal components and balance the multimodal joint training. With better

Table 2.1: Instruction tuning results for four MLLMs: BLIP-2, SALMONN, InternVL2-8B, and LLaVA-1.5-7B. These are pre-trained LLMs in vision and audio respectively. For questions-answering tasks like A-OKVQA, IconQA, TextVQA, and ClothoAQA, we report the accuracy score of the generated answers. For audio captioning tasks (MACS and SDD), we report the Rouge-L metric that compares the generated caption with candidate captions. We highlight the best method in bold font for each downstream task of instruction tuning.

Model	Task	Constant LR	Feature CD	Language CD	CoMMIT CLR	CoMMIT
BLIP-2	A-OKVQA	54.06	57.99	49.87	60.44	**64.37**
	IconQA	37.16	35.48	34.47	**39.09**	38.65
	TextVQA	26.48	18.00	19.44	27.66	**28.12**
SALMONN	ClothoAQA	42.49	45.80	38.52	**52.86**	50.55
	MACS	24.60	22.41	23.64	23.81	**25.06**
	SDD	15.10	5.70	**15.74**	15.07	15.33
InternVL2	A-OKVQA	76.59	73.19	79.47	78.00	**80.52**
	IconQA	80.94	**83.20**	81.60	80.85	82.87
	TextVQA	65.22	65.60	65.08	65.18	**67.00**
LLaVA-1.5	A-OKVQA	79.20	77.64	76.94	77.82	**79.55**
	IconQA	64.09	64.16	58.17	65.78	**69.60**
	TextVQA	41.98	43.34	**49.32**	47.80	49.30

Table 2.2: Comparison on A-OKVQA, TextVQA, and IconVQA with BLIP-2 backbone model. Baselines are Constant LR that direct fine tune the backbone model with various learning rate.

Method	A-OKVQA	TextVQA	IconVQA
LR=1e-5	50.30	27.58	35.45
LR=1e-4	54.06	26.48	37.16
LR=1e-3	45.24	20.60	34.93
CoMMIT	**64.37**	**28.12**	**38.65**

coordinated multimodal learning, CoMMIT-CLR and CoMMIT consistently improve Constant LR across modalities and downstream tasks. In addition, the proposed regularization in CoMMIT can promote larger step sizes in gradient descent, which enlarges differences in the generated output distributions between different time steps. This prevents learning from being stuck at local optima, which can be especially beneficial for modality-specific captioning tasks whose optimization space can be relatively larger than question-answering tasks.

Comparing various learning rates. In Table 2.2, we compare CoMMIT with results of constant LR with different learning rates. We report the results on tasks of A-OKVQA, TextVQA, and Icon-

VQA with BLIP-2 backbone model. We can observe that our proposed CoMMIT outperforms the Constant LR baselines by a significant margin. In addition, we can also find that no fixed value of learning rate consistently yields the best performance for Constant LR, while our proposed CoMMIT can dynamically adjust its learning rate. These results demonstrate the necessity and effectiveness of dynamic learning rate adjustment for balanced learning between the feature encoder and LLM in multimodal instruction tuning.

2.8 Theoretical Analysis

In this section, we present the computation and proof of a new convergence bound with our proposed method CoMMIT. Our theoretical analysis demonstrates that it achieves a faster convergence rate compared to the imbalanced MLLM instruction tuning.

2.8.1 Setup and Notations

Consider a non-convex random objective function $F : \mathbb{R}^d \to \mathbb{R}$ expressed as the average of K component functions, $F(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x)$, where each $f_k(x)$ is an i.i.d sample. Our goal is to minimize $\mathbb{E}F(x)$ over $x \in \mathbb{R}^d$. We also define $\mathbb{E}_{k-1}$ as the conditional expectation with respect to $f_1, f_2, \cdots, f_k$. Following the notation in Adam (Kingma and Ba, 2015), let m_k, v_k, $x_k \in \mathbb{R}^d$ be vectors at iteration k. The i-th component of each vector is denoted by $m_{k,i}$, $v_{k,i}$, $x_{k,i}$. Building upon the insight of Défossez et al. (Défossez et al., 2020) regarding the presence of two bias correction terms m_k and v_k, we define $\alpha_{k,i} = \alpha_i\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$. Notably, we opt to drop the correction term for m_k due to its faster convergence compared to v_k.

Aligned with our proposed methodology, we incorporate two additional terms into the original Adam algorithm. To prevent learning from oscillating too heavily toward either S or X, we adapt λ according to the moving average $\tilde{\kappa}$. To mitigate the risk of vanishing or exploding gradients, we introduce $h(x)$ as an extra term in the gradient updates defined in Section 2.6 to enhance training stability and support the overall robustness of the learning process. We fix $\beta_1 = 0$, $0 < \beta_2 \leq 1$,

$\alpha_{k,i} > 0$, $\epsilon = 10^{-8}$, and initialize $m_0 = 0$ and $v_0 = 0$. The updated rules follow,

$$v_{k,i} = \beta_2 v_{k-1,i} + (\lambda \nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1}))^2 \tag{2.12}$$

$$x_{k,i} = x_{k-1,i} - \alpha_k \frac{\lambda \nabla_i f_k(x_{k-1}) + \nabla_i h_k(x_{k-1})}{\sqrt{v_{k,i} + \epsilon}} \tag{2.13}$$

Throughout the proof, we assume the norm of the gradients $\|\nabla f(x) + \nabla h(x)\|$ is bounded by $R - \sqrt{\epsilon}$. The small constant ϵ is used for numerical stability.

2.8.2 Necessary Assumptions

We state the necessary assumptions (Bertsekas et al., 2003) commonly used when analyzing the convergence of stochastic algorithms for non-convex problems:

Assumption 2.5. *The minimum value of $f(x)$ is lower-bounded,*

$$\forall x \in \mathbb{R}^d, \ f^* = \min f(x).$$

Assumption 2.6. *The gradient of the non-convex objective function f is L-Liptchitz continuous (Nesterov, 2013). Then $\forall x, y \in \mathbb{R}^d$, the following inequality holds,*

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2} \|x - y\|_2^2.$$

2.8.3 Convergence Analysis

Following the second Theorem outlined by Défossez et al. (Défossez et al., 2020), we calculate the convergence bound of our algorithm with a dynamic learning rate and loss function.

Theorem 2.7. *Given the assumptions in the original paper and applying the descent lemma, for all components of the step sizes and gradients, update α_i with the corresponding value from H_t^S and H_t^T. Let $\{x_k\}$ be a sequence generated by the optimizer, with $0 < \beta_2 \leq 1$, and $\alpha_i > 0$. For*

any time step K, we have,

$$\mathbb{E}\|\nabla F(x_k)_2^2\| \leq 2R\frac{F(x_0) - f^*}{\lambda\alpha_i K} + C \tag{2.14}$$

where

$$C = \frac{1}{K}\left(\frac{2\alpha_i R}{\sqrt{1-\beta_2}} + \frac{\alpha_i^2 L}{2(1-\beta_2)}\right)\ln\left(\frac{(1-\beta_2^k)R^2}{(1-\beta_2)\epsilon\beta_2}\right)$$

The detailed proof is provided in the original paper.

CoMMIT adjusts both λ and $h(x)$ to balance multimodal learning progress. The parameter λ, which measures the balance between feature and language learning, remains greater than 1 during training, driven by the balance metric $\tilde{\kappa}$. By avoiding learning oscillations, λ can grow even larger, contributing to faster learning. When $\tilde{\kappa} > 1$, the model suffers from insufficient feature learning. CoMMIT reduces the learning rate of the LLM to balance learning, ensuring $\lambda = \frac{\tilde{\kappa}_t+1}{\tilde{\kappa}_{t-1}+1} > 1$. Conversely, when $\tilde{\kappa} < 1$, the model suffers from insufficient language learning, ensuring $\lambda = \frac{1/\tilde{\kappa}_t+1}{1/\tilde{\kappa}_{t-1}+1} > 1$. Notably, $\nabla h(x)$ is directly added to $\nabla f(x)$ to induce gradient changes, which further contributes to the increase of λ, resulting in a faster convergence rate.

In this section, we show the proof using Adam as the base optimizer. Due to the reason that CoMMIT does not modify the optimization algorithm itself, the theorem can also be extended to any gradient-based optimization method such as the stochastic gradient descent.

2.9 Chapter Summary

In this work, we address the challenge of imbalanced learning between the feature encoder and the backbone LMM during MLLM instruction tuning. Through theoretical analysis and empirical observations, we uncovered how this imbalance can lead to the dilemmas of oscillation and biased learning. To mitigate these problems, we proposed **CoMMIT**, a novel approach that dynamically coordinates the learning rates of the feature encoder and LLM backbone. Our **CoMMIT** also included regularization on the gradient gradients that promotes training sufficiency. Our theoreti-

cal and empirical analyses demonstrate that **CoMMIT** improves the balance between the training of the feature encoder and the LLM. Experiments across multiple vision and audio downstream instruction tuning tasks illustrate that the training with **CoMMIT** for MLLMs is more effective compared to baselines.

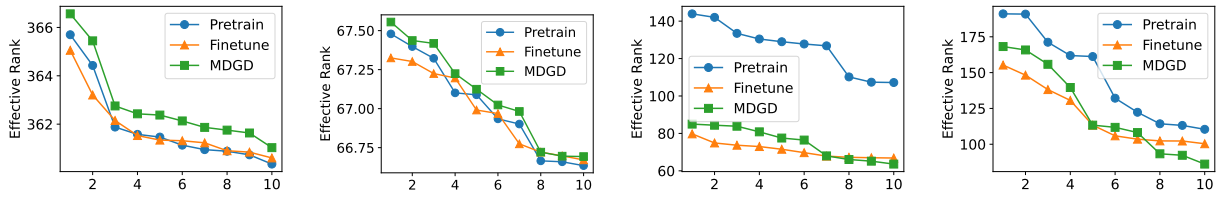
This chapter, in full, is a reprint of the material as it appears in "CoMMIT: Coordinated Instruction Tuning for Multimodal Large Language Models" by Xintong Li, Junda Wu, Tong Yu, Rui Wang, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Julian McAuley, and Jingbo Shang, published in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Modality-Decoupled Gradient Descent (MDGD)

3.1 Introduction

Multimodal large language models (MLLMs) enhanced visual understanding and reasoning by pre-training on large-scale multimodal datasets with comprehensive visual descriptions that integrate textual and visual knowledge Liu et al. (2023a); Yao et al. (2024b); Li et al. (2023b); Bai et al. (2023); Liu et al. (2024c); Wu et al. (2024f). These models achieve strong performance across various vision-language tasks, such as visual question answering Jin et al. (2024); Wu et al. (2025d), multimodal reasoning Zhang et al. (2024d); Jiang et al. (2024b); Yan et al. (2024), mul-



(a) LLaVA on OKVQA (b) LLaVA on POPE (c) MiniCPM on PathVQA (d) MiniCPM on POPE

Figure 3.1: The top-10 image tokens with the highest effective ranks on OKVQA and POPE encoded by LLaVA, and PathVQA and POPE encoded by MiniCPM. We compare pretrained, finetuned, and MDGD-finetuned models. Effective rank Wei et al. (2024) quantifies representation richness, which show that MDGD preserves higher effective rank, mitigating visual forgetting.

timodal recognition Shenoy et al. (2024); Wu et al. (2024f), and personalized multimodality Wu et al. (2024b, 2025a); Huang et al. (2025b). However, adapting pre-trained MLLMs to downstream tasks via instruction-tuning Li et al. (2025d, 2024a, 2023a); Panagopoulou et al. (2023); Liu et al. (2024c) presents a critical challenge of visual forgetting. Unlike pre-training, where models receive rich visual-text alignment, instruction-tuning is often text-driven with limited direct visual supervision. This shift in training focus leads to the degradation of pre-trained visual encoding Zhou et al. (2024a); Niu et al. (2024); Li et al. (2025d); Ko et al. (2023), negatively impacting model generalizability across downstream tasks that require strong visual knowledge Bai et al. (2024); Huang et al. (2024b). Addressing this challenge is essential for ensuring MLLMs retain their visual capabilities while aligning with new tasks efficiently.

While several approaches have attempted to mitigate catastrophic forgetting in neural networks through direct fine-tuning and continual learning methods Shi et al. (2024); Wu et al. (2024i); Zhu et al. (2024b); Zheng et al. (2024), these methods often overlook the unique challenge of preserving visual knowledge in multimodal large language models (MLLMs). Directly fine-tuning MLLMs on new tasks often leads to overfitting to textual instructions while inadvertently suppressing visual representations (Zhai et al., 2023). Existing continual learning strategies, such as regularization and replay methods, tend to focus on retaining language-based knowledge, neglecting the trade-off between compressing visual representations and aligning them with task-specific instructions Zhou et al. (2024a); Niu et al. (2024); Li et al. (2025d); Ko et al. (2023), leading to the degradation of pre-trained visual knowledge. Task-orthogonal gradient descent techniques have shown promise in disentangling gradients for multi-task optimization. However, their practical application in MLLMs poses unique challenges. MLLMs are pre-trained on vast and heterogeneous multimodal datasets Liu et al. (2023a); Li et al. (2023b); Bai et al. (2023), where it is challenging to isolate task-specific gradients, causing the components critical for visual understanding to become entangled with other features.

To gain a fundamental view of the challenge of visual knowledge forgetting in MLLM instruction tuning, we adopt an information bottleneck (IB) perspective that characterizes the trade-

off between retaining input information and ensuring output predictiveness Tishby et al. (2000). To investigate the degradation of crucial pre-trained visual knowledge, we introduce a novel perspective that leverages effective rank to quantify the richness of the encoded visual representation from MLLMs. Specifically, we illustrate the visual forgetting problem in Figure 3.1, where we observe a consistent effective rank reduction problem caused by MLLM instruction tuning. Based on this view, we propose a modality-decoupled gradient descent (MDGD) method, which disentangles the optimization of visual understanding from task-specific alignment, MDGD regulates gradient updates to maintain the effective rank of visual representations compared with pre-trained MLLMs, while mitigating the over-compression effects described by the information bottleneck. Intuitively, visual forgetting occurs due to the shift from rich multimodal pre-training to instruction-tuning, where text-based supervision dominates without direct visual supervision. By explicitly decoupling the task-specific alignment with visual representation learning, MDGD preserves expressive and robust visual features. To further improve efficiency in instruction-tuning, we introduce a memory-efficient fine-tuning strategy using gradient masking, which selectively updates a subset of model parameters for parameter-efficient fine-tuning (PEFT). This approach reduces computational overhead while ensuring that crucial pre-trained visual representations are retained.

We summarize our contributions as follows:

- We analyze the visual knowledge forgetting problem in MLLM instruction tuning and frame the problem through the lens of effective rank and information bottleneck theory.
- We propose MDGD, which decouples visual optimization from task-specific alignment to preserve visual representations and introduces a PEFT variant MDGD-GM to reduce computational overhead through gradient masking.
- We conduct comprehensive experiments on various MLLMs and downstream tasks, demonstrating that MDGD effectively mitigates visual forgetting while enabling strong adaptation to new tasks.

3.1.1 Visual Knowledge Forgetting in MLLMs

Catastrophic forgetting, where a model loses previous knowledge while learning new tasks, is a major challenge in continual learning Wang et al. (2023e). This problem is now widely recognized in LLMs and MLLMs Wu et al. (2024i); Luo et al. (2023); Zhai et al. (2023). Although various methods, including fine-tuning, task-orthogonal gradient descent, knowledge distillation, and replay, have been adapted to mitigate forgetting Shi et al. (2024); Wu et al. (2024i); Zhu et al. (2024b); Zheng et al. (2024), they often fail to preserve rich visual features. Fine-tuning on new tasks tends to overfit text and suppress visual information, while parameter-efficient methods like LoRA also suffer from forgetting Fawi (2024); Liu et al. (2024a). Model Tailor Zhu et al. (2024b) adapts the LLM backbone but does not address visual knowledge forgetting, which can lead to hallucination or degraded generalization Zhai et al. (2023). In contrast, our approach synchronizes the training of the visual encoder and LLM, preserving pre-trained visual knowledge during instruction tuning.

3.1.2 Information Theory in LLMs

The Information Bottleneck (IB) principle Tishby et al. (2000) has been used in LLMs to compress input while retaining task-relevant information Delétang et al. (2023); Valmeekam et al. (2023); Wei et al. (2024); Wu et al. (2022a). Prior works use IB to extract robust features Zhang et al. (2022); Wu et al. (2024e) and enable feature attribution Li et al. (2022); Jiang et al. (2020), but these focus on language models, not multimodal settings Yang et al. (2025). Existing information-theoretic transfer learning methods Tseng et al. (2024); Wu et al. (2024e); Ling et al. (2024) also do not address the unique challenges of MLLMs, where modalities are deeply entangled. Our method instead uses effective rank to measure and counteract visual representation compression. The proposed MDGD explicitly decouples visual learning from task alignment, going beyond previous IB-based approaches.

3.2 Preliminary

Task Definition. Given an MLLM π_θ and instruction-tuning dataset D, the image prompt $I \in \Omega$ is encoded by a visual encoder f into a sequence of M visual tokens $f(I) = X^v = (x_1^v, x_2^v, \ldots, x_M^v)$. During instruction tuning, the textual instructions $T \in D$ are tokenized as $X^l = (x_1^l, x_2^l, \ldots, x_N^l)$ using the tokenizer of the backbone LLM, which is querying the MLLM to generate textual responses conditioned on the multimodal inputs,

$$\hat{y}_k \sim \pi_\theta(\cdot \mid X^v, X^l, y_{<k}). \tag{3.1}$$

Therefore, the learning objective of visual instruction-tuning for K samples is to maximize the average log-likelihood of the ground truth answer tokens $y = (y_1, y_2, \ldots, y_T)$ of each sample,

$$\mathcal{L}_{vl}(\theta) = -\sum_{t=1}^{T} \log \pi_\theta \left(y_t \mid X^v, X^l, y_{<t}\right), \tag{3.2}$$

where multimodal instructions X^v and X^l both serve as generation conditions.

An Information Bottleneck Perspective on Visual Knowledge Forgetting. In multimodal models, the information bottleneck Mai et al. (2022) (IB) framework provides a powerful lens to understand how representations are formed. In our setting, the IB principle seeks a representation Z that is maximally informative about the output y while discarding irrelevant details from the inputs. For an MLLM that processes visual inputs X^v and textual inputs X^l, a full IB objective might take the form:

$$\min_{\theta} \quad \mathcal{L}_{\text{IB}}^{\text{vision}}(\theta) = -I(y; Z) + \beta\, I(X^v; Z). \tag{3.3}$$

where $I(\cdot;\cdot)$ denotes mutual information and β controls the trade-off between predictive power and compression. This formulation explicitly highlights the risk of discarding visual details when the model is optimized primarily to predict y. Based on the information bottleneck view, we further analyze the visual forgetting problem in the original paper.

Effective Rank as a Measure of Representation Richness. To quantify the information content retained in a representation, we use the effective rank metric Roy and Vetterli (2007). Given a representation matrix Z whose singular values are $\{\sigma_i\}$, the effective rank is defined as:

$$\text{erank}(Z) = \exp\left(-\sum_i p_i \log p_i\right), \tag{3.4}$$

where $p_i = \sigma_i / \sum_j \sigma_j$. This measure, based on the entropy of the singular value distribution, captures the "richness" or intrinsic dimensionality of Z. A higher effective rank indicates that the representation spans a larger subspace, whereas a lower effective rank implies that the representation has been overly compressed.

3.3 MDGD: Modality-Decoupled Gradient Regularization and Descent

Motivated by the visual forgetting problem caused by the degradation of multimodal encoding, we introduce a modality-decoupling gradient regularization (**MDGD**) to approximate orthogonal gradients between visual understanding drift and downstream task optimization. Specifically, leveraging modality-decoupled gradients $\bar{g}_\theta$ and $\bar{g}_\phi$ derived from the current MLLM and a pre-trained MLLM respectively, we propose a gradient regularization term $\tilde{g}_\theta$ for more efficient multimodal instruction tuning, which promotes the alignment of downstream tasks while mitigating visual forgetting Zhu et al. (2024b). Since MDGD requires the estimation of parameter gradients, we could not directly apply parameter-efficient fine-tuning methods (*e.g.*, LoRA Hu et al.). Thus, we alternatively formulate the regularization as a gradient mask $M_{\tilde{g}_\theta}$, which allows efficient fine-tuning only on a subset of masked model parameters.

3.3.1 Modality Decoupling

Based on the information bottleneck objective in Eq. (3.3), the objective encourages the model to maximize $I(y; Z)$ while compressing $I(X^v; Z)$ Tishby et al. (2000); Alemi et al. (2017). In practice, this compression may discard useful visual details, leading to visual forgetting. To mitigate such compression and preserve the pre-trained visual knowledge, we follow the KL divergence loss $D_{\text{KL}}\Big(\mu_\phi(X^v) \,\Big\|\, \pi_\theta(X^v)\Big)$ to constrain the current model's visual representation $\pi_\theta(X^v)$ to remain close to the pre-trained distribution $\mu_\phi(X^v)$, thereby preserving the mutual information $I(X^v; Z)$ that would otherwise be reduced by the compression Hinton (2015); Lopez et al. (2018). However, since MLLMs cannot directly track the distributions of image tokens, we instead introduce an auxiliary loss function

$$\mathcal{L}_v(\phi, \theta) = \|\mu(X^v|\phi) - \pi(X^v|\theta)\|_1, \tag{3.5}$$

which approximates the KL divergence loss Zhu et al. (2022b, 2017) by penalizing discrepancies between the pre-trained visual representation and that obtained during instruction tuning.

In the MLLM instruction tuning, the visual output tokens (e.g., $\{z_k^{vl}\}_{k=1}^M$) are encoded as latent representations. Such visual encoding cannot be directly supervised by any learning objective but is learned through textual gradient propagation of the negative log-likelihood loss in downstream tasks. To approximate the visual optimization direction, we derive the gradients of $\mathcal{L}_v(\phi, \theta)$ for both the pre-trained MLLM π_ϕ and the current MLLM π_θ:

$$\begin{aligned} h_\phi &= \nabla_\phi \mathcal{L}_v(\phi) = \boldsymbol{\lambda}(\phi, \theta) \cdot \nabla_\phi \mu(X^v|\phi), \\ h_\theta &= \nabla_\theta \mathcal{L}_v(\theta) = -\boldsymbol{\lambda}(\phi, \theta) \cdot \nabla_\theta \pi(X^v|\theta), \end{aligned}$$

where $\boldsymbol{\lambda}(\phi, \theta) = \text{sign}\left(\mu(X^v|\phi) - \pi(X^v|\theta)\right)$. Intuitively, when the MLLM's visual understanding

drift causes visual forgetting, we further derive the orthogonal task gradients $\bar{g}_\phi$ and $\bar{g}_\theta$:

$$\bar{g}_\phi = \nabla_\phi \mathcal{L}_{vl}(\phi) - \frac{\nabla_\phi \mathcal{L}_{vl}(\phi)^\top h_\phi}{\|h_\phi\|^2} \cdot h_\phi, \tag{3.6}$$

$$\bar{g}_\theta = \nabla_\theta \mathcal{L}_{vl}(\theta) - \frac{\nabla_\theta \mathcal{L}_{vl}(\theta)^\top h_\theta}{\|h_\theta\|^2} \cdot h_\theta, \tag{3.7}$$

which enables **modality decoupling** of the downstream task loss gradient in Eq.(3.2) orthogonal to the visual understanding drift for the pretrained MLLM $\bar{g}_\phi \perp h_\phi$ and current MLLM $\bar{g}_\theta \perp h_\theta$.

Algorithm 1 MDGD: Modality Decoupled Gradients Descent

Inputs: Pre-trained MLLM μ_ϕ, current MLLM π_θ, instruction-tuning dataset D, and learning rate η
Outputs: The optimized model weights of π_θ
Initialize $\pi_\theta \leftarrow \mu_\phi$
for Receive minibatch $D_i \subset D$ **do**
    Calculate $\mathcal{L}_{vl}(\phi)$ of μ_ϕ, based on Eq.(3.2);
    Calculate $\mathcal{L}_{vl}(\theta)$ of π_θ, based on Eq.(3.2);
    Extract visual encodings of $\mu(X^v|\phi)$;
    Extract visual encodings of $\pi(X^v|\theta)$;
    Calculate $\mathcal{L}_v(\phi, \theta)$, based on Eq.(3.5);
    Derive orthogonal task gradients $\bar{g}_\phi$ and $\bar{g}_\theta$, according to Eq.(3.6);
    if Parameter-efficient fine-tuning **then**
        Calculate $M_{\tilde{g}_\theta}$,based on Eq.(3.10);
        Update the model following Eq.(3.11).
    else
        Calculate $\tilde{g}_\theta$, based on Eq.(5.12);
        Update the model following Eq.(3.9).
    end if
end for

3.3.2 Regularized Gradient Descent

The auxiliary loss in Eq. (3.5) preserves the visual representation at a distribution level via the feature alignment auxiliary loss in Eq. (3.5). However, the information bottleneck framework indicates that the gradient component compressing $I(X^v; Z)$ (*i.e.*, $\nabla_\theta I(X^v; Z)$), can harm visual preservation by reducing the effective rank of the features Achille and Soatto (2018); Lee et al. (2021).

To address this compression-induced drift, we incorporate an orthogonal gradient as a regularize. Motivated by multi-task orthogonal gradient optimization Yu et al. (2020); Zhu et al. (2022a); Dong et al. (2022), we leverage the gradient $\bar{g}_\phi$ from the pre-trained model μ_ϕ, which reflects the accumulated visual drift and approximates a global orthogonal learning effect in the downstream task. We then project the current model's gradient onto this direction:

$$\tilde{g}_\theta = \frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\|^2} \cdot \bar{g}_\phi. \tag{3.8}$$

In addition, to prevent discrepancies between the regularization and task gradients, we include the feature alignment auxiliary loss (Eq. (3.5)) in the overall objective. The final parameter update is:

$$\pi_\theta \leftarrow \pi_\theta - \nabla_\theta \mathcal{L}_{vl}(\theta) - \nabla_\theta \mathcal{L}_v(\theta) - \tilde{g}_\theta. \tag{3.9}$$

3.3.3 Enabling Parameter-efficient Fine-tuning of MDGD via Gradient Masking

Parameter-efficient fine-tuning (PEFT) methods, such as adapters Houlsby et al. (2019) and LoRA Hu et al., aim to reduce the computational cost and memory usage when fine-tuning models on downstream tasks under practical constraints Han et al. (2024). However, due to the requirement of directly estimating gradient directions on the pre-trained model parameters, MDGD cannot be directly applied to these PEFT methods, which introduce additional model parameters whose gradients are separate from the original model weights.

To address this challenge, we propose a variant, MDGD-GM, by formulating the gradient regularization term in Eq. (5.12) as gradient masking that selects model weights with efficient gradient directions. Specifically, we define it as

$$M_{\tilde{g}_\theta} = \mathbf{1}\left\{ \frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\| \|\bar{g}_\theta\|} \geq T_\alpha \right\}, \tag{3.10}$$

where T_α is determined by a percentile α of trainable parameters with the highest similarity scores between $\bar{g}_\theta$ and $\bar{g}_\phi$. Consequently, the optimization in Eq. (3.9) is reformulated as

$$\pi_\theta \leftarrow \pi_\theta - M_{\tilde{g}_\theta} \cdot (\nabla_\theta \mathcal{L}_{vl}(\theta) + \nabla_\theta \mathcal{L}_v(\theta)) \,. \tag{3.11}$$

We summarize and illustrate the optimization process of MDGD and MDGD-GM in Algorithm 1.

Table 3.1: Performance on various pre-trained tasks of LLaVA-1.5 models fine-tuned on Flickr30K and OKVQA. We report the best performance for each task in a **bold font** while the second best performance underlined.

Method	#Params	Pre-trained tasks						Target task	Metrics	
		GQA	**VizWiz**	**SQA**	**TextVQA**	**POPE**	**MMBench**	**Flickr30k**	**Avg**	**Hscore**
Zero-shot	–	61.94	50.00	66.80	58.27	85.90	64.30	3.5	55.82	59.86
Fine-tune	1.2B	56.26	44.45	28.34	38.98	38.40	50.56	**78.82**	47.97	45.26
LoRA	29M	17.74	40.63	5.38	30.48	2.40	9.55	64.18	24.33	20.49
Model Tailor	273M	52.49	42.28	67.15	43.89	82.88	63.40	75.40	61.07	59.85
MDGD	1.2B	67.71	48.18	**69.05**	57.32	**85.12**	65.43	73.47	**66.61**	**66.03**
w/o visual align	1.2B	57.64	36.95	53.96	32.84	30.43	56.66	65.58	47.72	46.19
MDGD-GM	124M	**69.89**	**51.22**	65.87	**58.18**	84.39	**66.42**	64.18	65.74	65.86

Method	#Params	Pre-trained tasks						Target task	Metrics	
		GQA	**VizWiz**	**SQA**	**TextVQA**	**POPE**	**MMBench**	**OKVQA**	**Avg**	**Hscore**
Zero-shot	–	61.94	50.00	66.80	58.27	85.90	64.30	0.14	55.34	59.58
Fine-tune	1.2B	62.98	40.59	59.84	48.38	71.42	51.98	69.10	57.76	56.79
LoRA	29M	63.44	41.61	51.29	48.02	75.27	37.31	**71.46**	55.49	54.12
Model Tailor	273M	60.39	**46.49**	**69.51**	**54.88**	**85.44**	63.32	38.10	59.73	61.48
MDGD	1.2B	**66.55**	42.72	64.60	52.54	85.17	61.73	62.29	62.23	62.22
w/o visual align	1.2B	66.39	39.89	60.19	52.40	84.92	62.97	62.39	61.31	61.22
MDGD-GM	124M	66.02	43.97	67.91	52.80	84.70	**63.97**	61.04	**62.92**	**63.07**

3.4 Experiments

Datasets To evaluate catastrophic forgetting, we follow the setup of Zhu et al. (2024b) and consider two models: LLaVA-1.5 (7B) and MiniCPM-V-2.0 (2.8B). For each model, we use a set of pre-trained tasks (including VQAv2, GQA, VizWiz, TextVQA, POPE, and MM-Bench) and fine-tuning tasks on previously unseen datasets (such as Flickr30k, OKVQA, TextCaps, and PathVQA). Full details on dataset composition are provided in the appendix.

Baselines We compare our approach against several baselines: standard fine-tuning following Zhu et al. (2024b), LoRA-based fine-tuning (Hu et al.), and Model Tailor (Zhu et al., 2024b). For Model Tailor, we report the original results from their paper for comparison.

Implementation Details All experiments use the official Huggingface implementations of LLaVA-1.5 and MiniCPM-V-2.0, with LoRA adapters where applicable. Models are fine-tuned using BFloat16 precision on 2 NVIDIA A100 80GB GPUs. We include fine-grained implementation details in the original paper.

3.4.1 Comparison Results

LLaVA-1.5 adapts better to downstream tasks but is more prone to visual forgetting. We study the visual forgetting problem on the LLaVA-1.5 MLLM, and report performance comparison results in Table 3.1. We observe that the pre-trained LLaVA enables efficient instruction tuning on target tasks, where the zero-shot performance is near zero. When the model is fine-tuned on the image caption task, Flickr30K, which largely differs from the pre-trained tasks of visual question-answering, the model can learn a degraded multimodal representation, which causes visual forgetting in its projected visual representation space. Fine-tuning on visual question-answering task OKVQA, which is similar to the pre-trained tasks, can also lead to MLLM's visual understanding drift, due to the limited image-text pairs existing in the downstream task.

MiniCPM-V-2.0 also experiences visual forgetting while limited in downstream task improvements. To validate the observation on a smaller MLLM, we report the comparison results of MiniCPM-V-2.0 with 2.8B model parameters in Table 3.2. We observe that compared with the LLaVA MLLM, MiniCPM suffers from less prominent visual forgetting. We attribute this observation to MiniCPM learning a more compact and constrained visual representation space during pre-training, causing the visual representations of target task images to be less aligned with those of the pre-trained MLLM. Consequently, MiniCPM exhibits limited improvement in downstream tasks, as its restricted ability to acquire additional visual knowledge leads to ineffective instruction tuning.

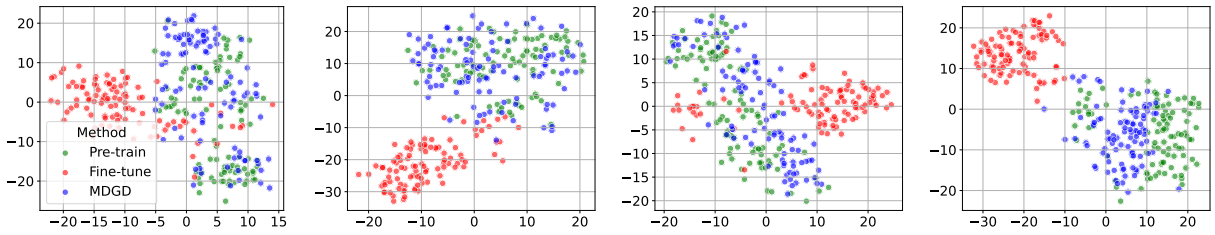
Table 3.2: Performance on various pre-trained tasks of MiniCPM-V2.5 models fine-tuned on PathVQA and TextCaps. We report the best performance for each task in a **bold font** while the second best performance underlined.

Method	#Params	Pre-trained tasks							Target task	Metrics	
		VizWiz	**A-OKVQA**	**OKVQA**	**TextVQA**	**IconQA**	**POPE**	**MMBench**	**PathVQA**	**Avg**	**Hscore**
Zero-shot	-	55.27	79.39	64.86	77.98	79.01	88.93	70.98	5.44	65.23	10.04
Fine-tune	517M	52.91	76.94	59.06	58.34	76.96	**89.60**	70.16	11.04	61.88	18.74
LoRA	35M	52.95	76.24	**64.45**	77.18	77.80	88.08	67.47	**15.03**	64.90	**24.41**
MDGD	517M	**55.73**	78.25	64.33	77.54	**79.45**	89.19	**71.94**	9.09	**65.69**	15.97
w/o visual align	517M	54.92	78.52	64.17	77.42	79.37	89.10	70.96	8.49	65.37	15.03
MDGD-GM	52M	55.04	**78.78**	64.31	**77.78**	79.10	88.76	70.98	5.72	65.06	10.52

Method	#Params	Pre-trained tasks							Target task	Metrics	
		VizWiz	**A-OKVQA**	**OKVQA**	**TextVQA**	**IconQA**	**POPE**	**MMBench**	**TextCaps**	**Avg**	**Hscore**
Zero-shot	-	55.27	79.39	64.86	77.98	79.01	88.93	70.98	15.77	66.52	25.50
Fine-tune	517M	52.03	77.73	59.16	67.24	78.67	88.20	71.42	**33.85**	66.04	**44.76**
LoRA	35M	53.30	78.17	63.99	77.68	78.28	87.31	69.23	32.41	67.55	43.80
MDGD	517M	**55.17**	78.17	63.67	76.08	79.40	**89.11**	71.58	28.90	67.76	40.52
w/o visual align	517M	51.35	78.08	63.06	76.48	78.99	88.98	71.30	25.93	66.77	37.35
MDGD-GM	52M	55.04	**78.43**	**65.26**	**78.08**	**79.65**	88.93	**71.88**	29.14	**68.30**	40.85

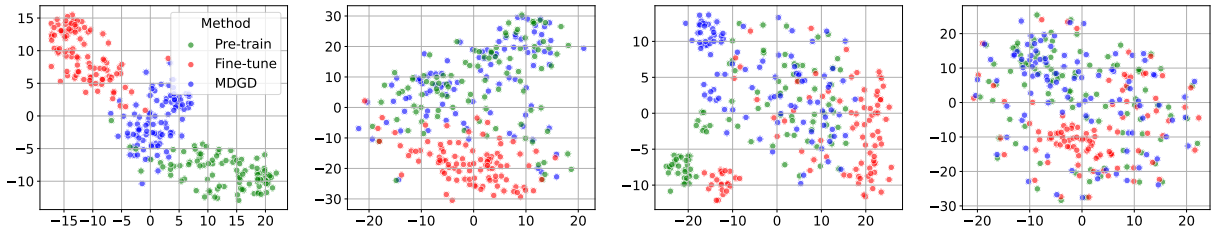
MDGD prevents visual forgetting while maintaining downstream task improvements. By employing MDGD in MLLM instruction tuning, we observe the LLaVA's average performance drops on pre-trained tasks when fine-tuned on OKVQA and also improves when fine-tuned on Flickr30K, which demonstrates the efficiency of MDGD in mitigating visual forgetting. For the smaller MLLM, MiniCPM, MDGD achieves comparable fine-tuning improvements with direct fine-tuning, while completely eliminating visual forgetting in the pre-trained tasks. MDGD and its variants consistently achieve the best average performance for both MLLMs, demonstrating its great potential for incremental learning on individual downstream tasks.

Comparison with baseline methods. Table 3.1 shows that MDGD consistently outperforms both LoRA fine-tuning and Model Tailor Zhu et al. (2024b) on LLaVA-1.5. LoRA suffers from visual forgetting due to projecting multimodal features into lower-rank spaces, especially on Flickr30K and OKVQA. Model Tailor, while effective for anti-forgetting in LLMs, is less robust for MLLMs and remains sensitive to the target dataset, performing better on Flickr30K than OKVQA. In contrast, MDGD achieves higher average scores and H-scores across datasets. In Table 3.2, MDGD improves average performance on MiniCPM tasks, reducing visual forgetting by 2.43% and 1.83% on PathVQA and TextCaps, respectively.



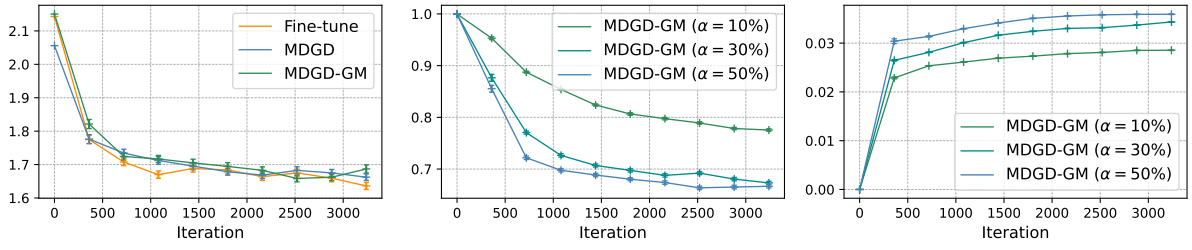
(a) z^{vl} on **PathVQA** (b) $\pi(X^v)$ on **PathVQA** (c) z^{vl} on **TextCaps** (d) $\pi(X^v)$ on **TextCaps**

Figure 3.2: T-SNE plots of the distribution of extracted visual $\pi(X^v)$ and multimodal z^{vl} representations from pre-trained MiniCPM, and models with direct fine-tuning and MDGD on PathVQA and TextCaps.



(a) z^{vl} on **OKVQA** (b) $\pi(X^v)$ on **OKVQA** (c) z^{vl} on **Fllickr30K** (d) $\pi(X^v)$ on **Fllickr30K**

Figure 3.3: T-SNE plots of the distribution of extracted visual $\pi(X^v)$ and multimodal z^{vl} representations from pre-trained LLaVA-1.5, and models with direct fine-tuning and MDGD on OKVQA and Flickr30K.



(a) Task Loss $\mathcal{L}_{vl}(\phi, \theta)$ (b) Cosine Similarity $\bar{g}_\theta^\top \bar{g}_\phi / \|\bar{g}_\phi\| \|\bar{g}_\theta\|$ (c) Representation Loss $\mathcal{L}_v(\phi, \theta)$

Figure 3.4: Illustration of (a) the learning process of three methods based on task loss $\mathcal{L}_{vl}(\phi, \theta)$, (b) the average regularized cosine similarity $\frac{\bar{g}_\theta^\top \bar{g}_\phi}{\|\bar{g}_\phi\| \|\bar{g}_\theta\|}$ in Eq.(3.10) for gradient masking at varying ratios, and (c) the visual representation loss $\mathcal{L}_v(\phi, \theta)$ in Eq.(3.5) for gradient masking at varying ratios α.

3.4.2 Ablation Study

Ablation study on visual alignment. We compare MDGD with its two variants, MDGD w/o visual align and MDGD-GM. MDGD w/o visual align enables MDGD without including visual representation loss $\mathcal{L}_v(\phi, \theta)$ Eq.(3.5), to understand the effect of directly optimizing to reduce the visual representation discrepancy between the current model and pre-trained model. We observe that MDGD w/o visual align maintains relatively comparable performance to MDGD on OKVQA and PathQA, due to the reduced need for visual representation adaptation in such visual question-answering tasks. In contrast, tasks like image captioning on Flickr30K and TextCaps benefit from feature alignment regularization, which directly mitigates visual understanding drift in the MLLM.

Ablation study on gradient masking. The other variant, MDGD-GM, leverages gradient masking to enable parameter-efficient fine-tuning (PEFT). We observe the PEFT variant of MDGD consistently achieves comparable performance across all tasks and backbone MLLMs, which only fine-tunes a subset of 10% original MLLM parameters used for direct fine-tuning and original MDGD. Different from conventional PEFT methods such as adapters, MDGD and its variants do not introduce additional parameters to the original model architecture, enabling incremental learning in an online setting (Maltoni and Lomonaco, 2019; Gao et al., 2023).

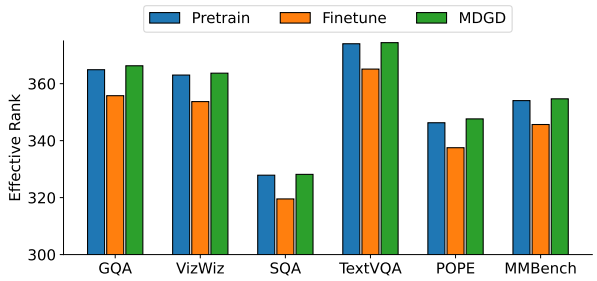
3.4.3 Representation Learning Analysis

T-SNE Analysis on Visual Representation To analyze the learning of visual and multimodal representation distributions in MLLMs, we create T-SNE Van der Maaten and Hinton (2008) plots to visualize the feature distributions extracted from pre-trained MLLMs, as well as MLLMs after standard fine-tuning and MDGD We illustrate the distributions of the multimodal features z^{vl} extracted from the last token of the multimodal instruction tokens, and the visual features $\pi_\theta(X^v)$ extracted from the last token of the input image tokens. We observe a consistent visual understanding drift in the MLLMs' visual representation spaces after standard fine-tuning on PathVQA and TextCaps with MiniCPM (Figure 3.2b and 3.2d). By employing MDGD to mitigate visual forgetting, we

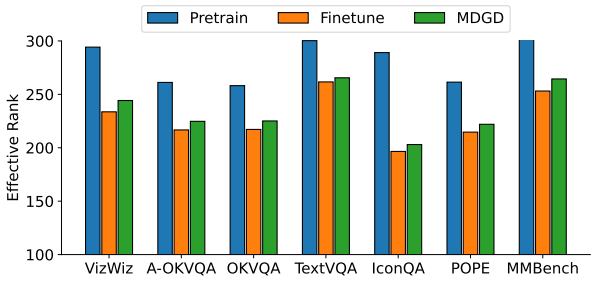
observe that visual understanding drift is effectively reduced, allowing the fine-tuned MLLM to retain pre-trained visual capabilities.

We further observe a distributional discrepancy in the multimodal representation z^{vl} of LLaVA (Figures 3.3a and 3.3c) between MDGD and the pre-trained MLLM. This discrepancy arises from the alignment of the MLLM to the target task through multimodal instructions, demonstrating effective adaptation to the downstream task of the LLaVA model. In addition, we also observe such multimodal distribution discrepancy reduces in a smaller MLLM, MiniCPM. This observation aligns with our findings on MiniCPM in Section 3.4.1, where we noted limited effects in model adaptation to downstream tasks. However, applying MDGD to MiniCPM mitigates visual forgetting by preventing degradation of both image and multimodal encodings into lower-rank representation spaces.

Effective Rank Analysis on Visual Representation To quantitatively analyze the visual forgetting problem, we calculate effective ranks of the visual representations extracted from the last hidden layer on the position of image tokens in individual MLLMs. We show the comparison results of LLaVA models in Figure 3.5a and MiniCPM models in Figure 3.5b. We observe that with both the backbone models of LLaVA and MiniCPM, directly fine-tuning the pre-trained models on downstream tasks can lead to a consistent reduction of effective ranks in visual representations. Such observations validate the hypothesis regarding the potential visual forgetting problem in MLLM instruction tuning. In addition, we can observe that MDGD achieves consistent improvements in effective ranks compared with the standard fine-tuning method for both backbone MLLMs across various pre-trained tasks. In Figure 3.5a, we observe that MDGD achieves comparable or even better effective ranks on pre-trained tasks, compared with the pre-trained LLaVA model. However, MDGD on MiniCPM in Figure 3.5b also suffers from the visual representation degradation problem, while MDGD consistently alleviates the problem. Such observation suggests a higher risk of visual forgetting in smaller-scale MLLMs.



(a) LLaVA models pretrained, finetuned, and fine-tuned with MDGD



(b) MiniCPM models pretrained, finetuned, and fine-tuned with MDGD

Figure 3.5: The effective rank comparison on individual downstream fine-tuning datasets.

3.4.4 Sensitivity Study

We evaluate the learning curves of MDGD and MDGD-GM compared with standard fine-tuning in Figure 3.4(a), where we observe that MDGD and MDGD-GM achieve comparable training efficiency compared with the standard fine-tuning method. We also investigate the sensitivity of gradient cosine similarity between $\bar{g}_\theta$ and $\bar{g}_\phi$ in Figure 3.4(b) and the representation loss in Figure 3.4(c), with respect to the gradient masking ratio in MDGD-GM. In Figure 3.4(b), we observe that MDGD-GM with lower gradient masking ratios can better align the modality-decoupled learning gradients between the target model and the pre-trained model, while MDGD-GM maintains over 70% alignment with 50% gradient masking. In Figure 3.4(c), we show that MDGD-GM with 50% gradient masking still effectively alleviates the visual representation degradation problem by reducing the visual representation discrepancy $\mathcal{L}_v$, while learning with a more active gradient can achieve better alignment.

3.5 Chapter Summary

In this work, we addressed the challenge of visual forgetting in MLLMs during instruction tuning by introducing a novel modality-decoupled gradient descent (MDGD) approach. MDGD disentangles the gradient updates for visual representation learning from task-specific alignment, thereby preserving the effective rank of pre-trained visual features and mitigating the over-compression effects highlighted by the information bottleneck perspective. This decoupling enables MLLMs

to retain rich visual knowledge while adapting robustly to new downstream tasks. Furthermore, our gradient masking variant, MDGD-GM, enhances memory efficiency and optimizes parameter usage, making fine-tuning both practical and scalable. Extensive experiments across various downstream tasks and backbone models demonstrate that MDGD not only effectively prevents visual forgetting but also outperforms existing strategies in achieving balanced multimodal representation learning and task adaptation. Our findings underscore the importance of preserving visual representations during instruction-tuning and offer a viable solution for efficient and effective multimodal learning in real-world scenarios.

This chapter, in full, is a reprint of the material as it appears in "Mitigating Visual Knowledge Forgetting in MLLM Instruction-tuning via Modality-decoupled Gradient Descent" by Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Lina Yao, Jingbo Shang, and Julian McAuley, published in *Findings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Offline Chain-of-Thought Evaluation and Alignment via Knowledge Graph Exploration (OCEAN)

4.1 Introduction

Offline policy evaluation aims to estimate a target policy model's performance with only collected data, without requiring direct interactions between the target policy and realistic environments. Previous offline evaluation methods focus on decision-making policies in recommender systems (Li et al., 2011), healthcare (Bang and Robins, 2005), and other scenarios where online experimentation is costly (Thomas et al., 2015; Bhargava et al., 2024), risky, and impractical (Yu et al., 2021). Recent studies in LLMs leverage human feedback to align models' behaviors with human preferences in single-turn generation (Ouyang et al., 2022; Rafailov et al., 2023) and multi-step reasoning tasks (Joshi et al., 2024). Due to the high cost of deploying LLMs online and interacting with human feedback, Bhargava et al. (2024) further enables offline evaluation of LLMs from logged human feedback to align LLMs' response generation.

However, considering annotators may not be comprehensive in various types of knowledge backgrounds and the associated reasoning, human feedback on chain-of-thought reasoning (Joshi

et al., 2024) can be more challenging to collect. In addition, since chain-of-thought reasoning involves a sequential decision-making process, the volume of collected human feedback can be exponentially increased. Due to such challenges, conventional reinforcement learning from human feedback (RLHF) methods (Ouyang et al., 2022; Bai et al., 2022a) can suffer from training inefficient and scalability issues.

Motivated by recent works in using knowledge graphs as side information to enable prompt engineering (Wang et al., 2024e), self-correction (Zhao et al., 2023b; Wang et al., 2023d; Li et al., 2024c; Wu et al., 2024d), evaluating chain-of-thought (Nguyen et al., 2024), and model fine-tuning (Wang et al., 2024c; Tang et al., 2024), we propose leveraging knowledge graphs as weak yet controllable knowledge reasoners to effectively measure the alignment between LLMs' multi-step chain-of-thought reasoning and multi-hop knowledge graph trajectories by *inverse propensity scores* (IPS) (Joachims et al., 2017). In contrast to the existing chain-of-thought evaluation (Nguyen et al., 2024) method, which relies on accurate chain-of-thought grounding on specific knowledge graphs, we propose verbalizing the knowledge graph trajectories and developing a knowledge graph policy that serves as a verbal reasoning mechanism over the graphs. Therefore, we can bridge the heterogeneity between knowledge graph and LLM reasoning forms, and the verbalized knowledge graph policy can be generalized to be compatible with various LLMs.

To enable controllable chain-of-thought alignment in LLMs, we principally track LLMs' decision-making process in generating chain-of-thought reasoning steps, by formulating the process as a Markov Decision Process (MDP) whose goal is to reach the correct final answer with minimal knowledge exploration and exploitation. Then, we propose offline chain-of-thought evaluation and alignment, `OCEAN`, which evaluates the generated chain of thoughts from off-policy LLMs through collected offline data samples with feedback from a knowledge graph. The improved KG-IPS approach considers the effects of feedback from the knowledge graph policy that aligns the model's chain-of-thought generation and the behavior policy, which prevents model degeneration. We prove that the KG-IPS estimator provides an unbiased estimate of the target policy, with a lower bound for the variance, and establish confidence intervals using sub-Gaussian con-

centration inequalities. To enable direct optimization of LLM policies, we leverage the proposed KG-IPS policy evaluation approach for LLM fine-tuning by directly maximize estimated policy values through gradient descent. Then we empirically evaluate the optimized LLM policy on three types of chain-of-thought reasoning tasks, and demonstrate the effectiveness of the proposed policy optimization method. We also observe relative performance improvements across evaluation tasks, without affecting LLMs' generalizability or generation quality. We summarize our contributions as follows:

- We propose an offline evaluation framework, `OCEAN`, which bridges the heterogeneity between LLM and knowledge graph reasoning, for effective evaluations of chain-of-thought.
- With the evaluation framework, we further develop a direct policy optimization method which enables efficient alignment with automatic feedback from the knowledge graph.
- To facilitate the evaluation and optimization, we model the knowledge-graph preference and derive feedback by developing a policy which verbalizes knowledge-graph trajectories.
- We provide a theoretical analysis of the unbiasedness and establish a lower bound for the variance of our KG-IPS estimator.
- Through comprehensive experiments, we demonstrate `OCEAN`'s effectiveness in aligning LLMs' chain-of-thought reasoning through direct optimization of the estimated policy value. `OCEAN` also achieves better performance on various downstream tasks without affecting LLMs' generalizability.

4.2 Related Work

Offline Policy Evaluation Offline policy evaluation (OPE) is essential when online deploying learned policies is risky and impractical (Levine et al., 2020). OPE has been applied to various practical applications, including evaluating the recommender system's behavior with offline collected user feedback (Gilotte et al., 2018; Jeunen, 2019). Recent work (Gao et al., 2024) also develops an OPE estimator for LLM evaluation based on human feedback. Different from previous works, we study and formulate chain-of-thought generation in LLM as an MDP and use

knowledge graph reasoning as automatic feedback to develop a KG-IPS policy value estimator.

LLM Alignment Reinforcement Learning from Human Feedback (RLHF) has been the dominant approach, optimizing LLMs using human-annotated data to align model behavior with user preferences (Ouyang et al., 2022; Bai et al., 2022a). DPO (Rafailov et al., 2023) and RRHF (Yuan et al., 2023b) are proposed to reduce the training instability of RLHF. Wu et al. (2023c) utilizes varying densities of human feedback to offer fine-grained rewards for RL finetuning, and Sun et al. (2024a) focuses on aligning LLMs with reward models driven by human-defined principles. To address RLHF's limitations such as heavy reliance on human input, alternative approaches like Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022b; Lee et al., 2023; Liu et al., 2023c) and self-alignment methods (Sun et al., 2024b) have been proposed, using AI-generated feedback to scale and automate alignment. Despite advancements, a key challenge remains in aligning LLMs' internal knowledge with their reasoning, resulting in flawed reasoning even after factual errors are corrected. Our approach focuses on improving chain-of-thought alignment by modeling reasoning paths as an MDP and using KGs to ensure both factual accuracy and human-like reasoning.

Chain-of-thought Reasoning Chain-of-thought prompting has been widely applied to elicit the strong reasoning abilities of LLMs (Wei et al., 2022; Chu et al., 2023c; Xia et al., 2024). By decomposing a complex problem into a sequence of intermediate sub-tasks, LLMs are able to focus on important details and solve the step-by-step (Huang and Chang, 2023; Yu et al., 2023). Despite the remarkable performance improvements, recent studies have found that LLMs often generate unfaithful chain-of-thought reasoning paths that contain factually incorrect rationales (Turpin et al., 2023b; Lanham et al., 2023). To address this, a number of works leverage LLMs' self-evaluation abilities to verify and refine each reasoning step (Ling et al., 2023; Madaan et al., 2023). As the factual errors in the generated chain-of-thought may also be caused by the limited or outdated parametric knowledge of LLMs, recent methods incorporate external knowledge sources to further edit unfaithful content in the reasoning path (Zhao et al., 2023b; Wang et al., 2023d; Li et al., 2024c; Wang et al., 2024f,b). While these methods focus on knowledge augmentation and editing

through prompts, our method, in comparison, directly aligns LLM internal knowledge with faithful and factual chain-of-thought, which avoids potential knowledge conflicts between parametric and non-parametric knowledge when generating reasoning paths.

4.3 Preliminary

We first provide the formulation of chain-of-thought reasoning in LLMs as an MDP. Then we discuss conventional knowledge graph reasoning, as an alternative to free-form generation by verbalizing structured knowledge graph reasoning paths into natural language, which is more statistically controllable and generates faithful reasoning paths to the knowledge graph.

4.3.1 Problem Formulation: CHAIN-OF-THOUGHT AS AN MDP

Given the prompt instruction q, chain-of-thought reasoning process in a causal language model π_θ includes the generation of a trajectory of reasoning steps $\boldsymbol{c} = (c_1, c_2, \ldots, c_T)$, before the final answer prediction y,

$$c_t \sim \pi_\theta(\cdot|q) = \prod_{i=1}^{t-1} \pi_\theta(c_i|q, c_{<i}), \quad y \sim \pi_\theta(\cdot|q) = \pi_\theta(y|q, \boldsymbol{c}) \prod_{i=1}^{T} \pi_\theta(c_i|q, c_{<i}),$$

where each reasoning step c_t comprises a sequence of tokens and the number of reasoning step T is determined by the model's generation. Controllable chain-of-thought generation can be challenging due to its nature in autoregressive sequential sampling (Lin et al., 2020), which produces a high-dimensional action space in sampling a reasoning step $\pi_\theta(c_t|q)$ containing multiple tokens.

Chain-of-thought reasoning can be viewed as a Markov Decision Process (MDP) (Sutton, 2018): starting with the instruction prompt q, the LLM sequentially decides and generates the next-step reasoning path c_t that navigates until it arrives at a target final answer y. Given the LLM policy π_θ, at time step t, each **state** $s_t \in \mathcal{S}$ comprises of the instruction prompt q and previously generated reasoning paths $(c_i)_{i=0}^{t-1}$. The **action** space $\{1, \ldots, |\mathcal{V}|\}^{N_t}$ in LLMs is a sequence of N_t tokens sampled from an identical and finite vocabulary set $\mathcal{V}$. The LLM policy π_θ samples next-step thought

based on current state as $a_t \sim \pi_\theta(a_t|s_t)$, which is a sub-sequence in the reasoning path $a_t \in c_t$ and the surrounding context $c_t \backslash a_t$ is deterministically generated by LLMs. The **transition** in chain-of-thought is concatenating each reasoning path to the current state as $s_{t+1} = [s_t, c_t]$. Then the **reward** function is to evaluate each thought given the state as $r_t = r(s_t, c_t)$. Although such formulation of chain-of-thought enables direct LLM on-policy optimization via reinforcement learning, direct interaction with knowledge graphs to collect per-step reward in LLMs can be practically challenging and require a large effort of engineering due to the discrepancy between the unstructured generation of LLMs and structured knowledge graphs (Pan et al., 2024). Therefore, we propose to offline evaluate and optimize the target policy aligning with knowledge graph preference.

4.3.2 Verbalized Knowledge Graph Reasoning

In contrast to chain-of-thought reasoning, conventional knowledge graph reasoning methods (Lin et al., 2018; Saxena et al., 2020) sample a entity-relation pair (r_t, e_t) at step t from a subset of the graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ consisting of the outgoing edges of current edge e_{t-1},

$$(r_t, e_t) \in \{(r', e') | (e_{t-1}, r', e') \in \mathcal{G}\}, \tag{4.1}$$

where the transition feasibility of the entity e_{t-1} to all the outgoing edges is entirely determined by $\mathcal{G}$. The process of knowledge graph reasoning starts with a decomposed triple (e_0, r_1, e_1) of the instruction q, and produces a chain of triplets $\boldsymbol{h} = (e_0, r_1, e_1, \ldots, r_T, e_T)$ by sampling from a policy μ,

$$(r_t, e_t) \sim \mu\left((r_t, e_t) | e_0, r_1, e_1, \ldots, r_{t-1}, e_{t-1}\right), \tag{4.2}$$

where the goal of such knowledge graph exploration is to arrive at the correct answer entity at the end of the search step T. By knowledge graph exploration, we can collect a set of trajectories $\mathbb{H} = \{\boldsymbol{h}_k\}_{k=1}^K$, which are used to estimate a parametric probabilistic policy μ_ϕ as a proxy to model the preference of the knowledge graph.

To align the action space between the knowledge graph preference policy μ_ϕ and the target

policy π_θ, we leverage a small language model as the backbone of μ_ϕ and fine-tune the model on verbalized trajectories as natural language contexts. Inspired by existing efforts in verbalizing structured knowledge graphs into natural language query (Seyler et al., 2017) and context (Agarwal et al., 2020; Wang et al., 2022), we leverage the GPT-4 (Achiam et al., 2023) model f to verbalize each chain of triplets $\boldsymbol{h}$ into a chain-of-thoughts $\boldsymbol{c} = f(\boldsymbol{h})$. The verbalized knowledge-graph trajectories are used to model knowledge graph preference in Section 4.4.2.

4.4 OCEAN: Offline Chain-of-thought Evaluation and AlignmeNt

We propose an offline evaluation of the chain-of-thought generation process aligned with knowledge graph preference. The off-policy estimator can be used for policy optimization that aligns LLMs with more faithful reasoning paths from knowledge graphs (Lin et al., 2024). We develop a small language model as a behavior policy that models the knowledge graph preference.

4.4.1 Offline Evaluation and Optimization

One of the most broadly used offline evaluation approaches is *inverse propensity scores* (Ionides, 2008; Dudík et al., 2011), which has been used for LLM-based offline policy evaluation for various purposes (Bhargava et al., 2024; Dhawan et al., 2024). Given the offline logged chain-of-thought trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^N$, where $\tau_i = (s_t^{(i)}, c_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)})_{t=0}^{T_i}$, we propose a KG-IPS estimator considering two-folded weights of entity tokens in the knowledge graph preference policy μ_ϕ and of non-entity tokens in the base LLM policy π_0 ,

$$\hat{V}_{KG-IPS}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i|c_t^{(i)}|}\sum_{t=1}^{T_i}\sum_{e\in c_t}\frac{\pi_\theta(e|s_t^{(i)})}{\lambda(e|s_t^{(i)})}\log\pi_0(e|s_t^{(i)}), \tag{4.3}$$

where $\lambda(e|s_t^{(i)}) = \mathbf{1}\{e \in a_t^{(i)}\}\cdot\mu_\phi(e|s_t^{(i)}) + \mathbf{1}\{e \in c_t^{(i)}\backslash a_t^{(i)}\}\cdot\pi_0(e|s^{(i)})$. We follow (Zhang

et al., 2024b) to use the log-likelihood score of each token in the base policy π_0 as the reward function.

Establishing the unbiasedness of the KG-IPS estimator is essential for reliable policy evaluation (Jiang and Li, 2016; Bhargava et al., 2024). We formalize this in the following lemma; the proof is given in the original paper:

Lemma 4.1. *The KG-IPS estimator provides an unbiased estimate of the target policy π_θ.*

The standard IPS estimator is known to have high variances considering large behavior discrepancies ($\pi_\theta(e|s_t^{(i)})/\mu_\phi(e|s_t^{(i)})$) between the behavior policy μ_ϕ and the target policy π_θ. In addition, by separately weighting the entity and non-entity tokens with μ_ϕ and π_0 respectively, we avoid the increasing variance accumulated from the long chain-of-thought reasoning process and maintain the LLM's behaviors on non-entity tokens without model degeneration. To further formalize our approach and illustrate the variance inherent in the KG-IPS estimator, we present the following Lemma, which provides a lower bound on the variance,

Lemma 4.2. *The variance of the KG-IPS estimator is lower bounded by $\frac{M^2}{4n}$, where M denotes the maximum value of the weighted terms, and n is the number of samples. Given $V(\theta)$ is the true value function of the target policy π_θ, applying the concentration inequality for sub-Gaussian variables, the KG-IPS estimator satisfies the following confidence interval with probability at least $1-\delta$:*

$$\left|\hat{V}_{\textit{KG-IPS}}(\theta) - V(\theta)\right| \leq O\left(M\sqrt{\log(1/\delta)/n}\right).$$

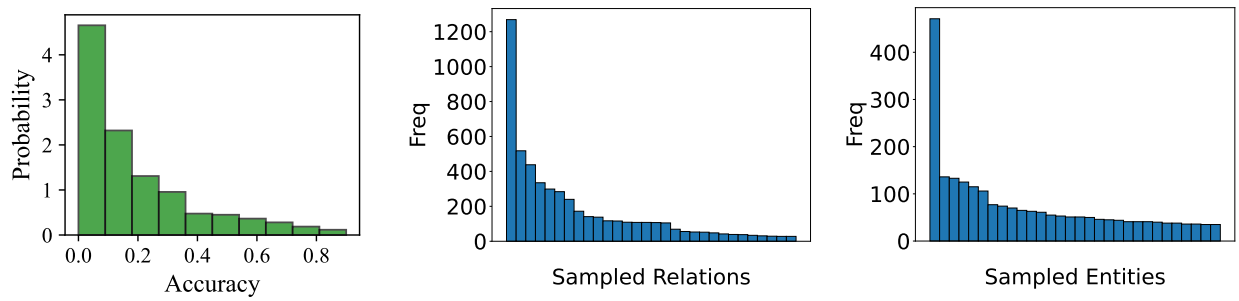
To further support our findings, we demonstrate that the optimal policy for the final reward is consistent with the optimal policy for the entity-based knowledge graph reward, which means the non-entity-based LLM reward can be considered as a regularization term that does not affect the optimal policy. See the original paper for a complete analysis.

In the end, we could directly optimize the target policy by maximizing the estimated value

function through policy gradient,

$$\theta \leftarrow \theta + \nabla_\theta \hat{V}_{KG-IPS}(\theta). \tag{4.4}$$

4.4.2 Knowledge Graph Preference Modeling



(a) Distribution of QA Accuracy (b) Distribution of Relations (c) Distribution of Entities

Figure 4.1: Sampling distributions of (a) trajectories in the knowledge graph that are verbalized as multi-step QA tasks and successfully answered by the LLM itself, (b) relations, and (c) entities in the knowledge graphs. We show their frequencies of the appearance in the trajectories sampled from the *Wikidata5M* (Wang et al., 2021) knowledge graph.

To facilitate the evaluation and optimization, we model knowledge graph preference and derive feedback by developing the behavior policy μ_ϕ which verbalizes knowledge-graph trajectories. Randomly sampled trajectories $\mathbb{H}$ from $\mathcal{G}$ in Section 4.3.2 contain samples that may not be transformed into a chain of thoughts leading to a reasonable question-answering. Following conventional self-consistent measurement (Wang et al.; Manakul et al., 2023), given a sampled trajectory $\boldsymbol{h}$ and its verbalized chain of thoughts $\boldsymbol{c}$, we prompt the GPT-4 model to propose a question q related to the first entity $e_0 \in \boldsymbol{h}$ whose answer should be exactly the last entity $e_T \in \boldsymbol{h}$, and query the GPT-4 model with the proposed question,

$$\hat{q} \sim f(q|e_0, e_T, \boldsymbol{c}), \quad \hat{y} \sim f(y|\hat{q}, \boldsymbol{c}), \quad R(\boldsymbol{h}|\boldsymbol{c}) = \mathrm{E}\left[\mathbf{1}\left\{e_T = \hat{y}\right\}\right],$$

where the reward of the trajectory is determined by the answer accuracy. In Figure 4.1a, we present the probability distribution of sampled trajectories, with respect to the number of correct answers

generated per trajectory from ten differently sampled questions associated with each trajectory. Based on such self-consistency measurement, we estimate the reward function $R(\boldsymbol{h}|\boldsymbol{c})$ as the normalized question-answering accuracy. Then we fine-tune the preference policy μ_ϕ directly via policy gradient optimization,

$$\nabla_\phi J(\phi) = \nabla_\phi \sum_{k=1}^{K} \sum_{t=0}^{|\boldsymbol{c}_k|-1} R(\boldsymbol{h}_k|\boldsymbol{c}_k) \log \mu_\phi(y_t|q, y_{<t}).$$

Based on the distribution of relations (Figure 4.1b) and entities (Figure 4.1c) in the sampled knowledge graph trajectories, we observe that the relation distribution is relatively more skewed toward the most frequent relations. This suggests that the verbalized knowledge graph reasoning policy is likely to focus on more frequent reasoning transitions, potentially enhancing its ability to learn meaningful patterns. In contrast, the entity distribution shows a relatively short tail, which may help mitigate the risk of overfitting to specific entities or knowledge biases.

4.5 Experiments

In this section, we evaluate our proposed method, `OCEAN`, by conducting chain-of-thought alignment on four LLM backbone models and evaluating several downstream tasks. We show our method's effectiveness in chain-of-thought alignment and its generalizability in various tasks to understand (i) whether the proposed optimization approach sufficiently aligns LLMs' chain-of-thought behaviors with higher estimated values on multi-hop question-answering tasks, (ii) how the proposed method performs on knowledge-intensive question-answering tasks and (iii) whether the post-alignment LLM generalizes on commonsense reasoning tasks.

4.5.1 Implementation Details

Datasets. Following prior work (Zhang et al.), we evaluate our approach across three key aspects of question answering. For *knowledge-intensive reasoning*, we use datasets requiring deeper domain understanding. The **AI2 Reasoning Challenge (ARC)** (Clark et al., 2018) tests

models' advanced reasoning abilities with grade-school science questions. **PubMedQA** (Jin et al., 2019) assesses biomedical reasoning by requiring yes/no/maybe answers from research abstracts. Finally, **SciQA** (Auer et al., 2023) challenges models to reason over scientific knowledge using the Open Research Knowledge Graph (ORKG).

The second focus is *multi-hop reasoning*, where models must combine information from multiple sources. **HotpotQA** (Yang et al., 2018) requires reasoning across multiple Wikipedia documents. **MuSiQue** (Trivedi et al., 2022b) includes complex questions needing 2-4 inference hops. **StrategyQA** (Geva et al., 2021) tests implicit reasoning, where steps must be inferred. These datasets assess the models' ability to perform connected reasoning tasks for complex questions.

Lastly, we focus on *commonsense reasoning*, crucial for understanding everyday knowledge. We evaluate three different commonsenseQA benchmarks, **CSQA** (Talmor et al., 2021), **CSQA2** (Saha et al., 2018), and CSQA-COT1000 (Li et al., 2024b). In addition, **OpenBookQA** (Mihaylov et al., 2018) tests models' ability to combine common knowledge with science facts for elementary-level questions. **WinoGrande** (Sakaguchi et al., 2021) challenges models to avoid biases in commonsense reasoning. These tasks evaluate a model's general abilities in commonsense question-answering.

Baselines. We experiment with four backbone LLMs: Gemma-2 (Team, 2024a) with 2B model parameters, Llama-3 (AI@Meta, 2024) with 8B model parameters, Phi-3.5-mini (Abdin et al., 2024) with 3.8B model parameters, and Mistral-0.2 (Jiang et al., 2023) with 7B model parameters. We use the instruction fine-tuned version of backbone LLMs for better instruction following abilities in question-answering. For chain-of-thought alignment in `OCEAN`, we use the CWQ question-answering dataset (Talmor and Berant, 2018) as the source data, in which the question-answering pairs are developed from knowledge graphs. `OCEAN` only uses CWQ questions for the LLM to generate chain-of-thought reasoning paths, which are further aligned using the knowledge graph preference model, without directly supervised learning on the ground-truth answers. To compare with direct supervised learning, we also enable instruction-tuning as a baseline (SFT), which is fine-tuned with the question as instruction and the answer as the response.

Knowledge Graph Preference Model. The knowledge graph preference model is developed based on the pre-trained GPT2-Medium model (Radford et al., 2019). We collected 6K question-answering pairs from the Wikidata5M (Wang et al., 2021) knowledge graph based on the sampling strategy in Section 4.4.2. The sampled knowledge graph trajectories are composed into natural language prefixed by the corresponding questions by the GPT-4 model, which verbalizes the knowledge graph reasoning trajectories and aligns with generative language models' behaviors. The model is then fine-tuned with a base learning rate of $1e-4$ for 10 epochs with a linear learning scheduler.

4.5.2 Multi-hop Question Answering

We evaluate the chain-of-thought reasoning performance of `OCEAN` compared with base LLMs and supervised fine-tuning (SFT), in three multi-hop question-answering tasks in Table 4.1. Comparing SFT and Base LLMs, we observe similar knowledge inconsistency as in knowledge-intensive tasks. Although SFT improves on MuSiQue with Gemma-2 and Mistral-0.2 backbones whose base models' performance is relatively inferior on this task, such knowledge-inconsistent problems result in worse performance on other downstream tasks.

Since `OCEAN` is aligned to incorporate more knowledge-faithful chain-of-thought reasoning patterns learned from knowledge graph reasoning policy without directly editing its internal knowledge, `OCEAN` maintains its generalizability in adapting to downstream tasks. We observe that `OCEAN` consistently improves on the policy estimated value $\hat{V}(\theta)$ through direct policy optimization proposed in (5.2), which demonstrates the effectiveness of the developed optimization method. Regarding the question-answering accuracy, `OCEAN` improves base LLMs, which achieves the best performance on HotpotQA and StrategyQA without context.

4.5.3 Knowledge-intensive Question Answering

To understand the effectiveness of `OCEAN` in knowledge-intensive question-answering tasks, we show performance comparison with base LLMs (`Base`) and supervised fine-tuning (`SFT`) in Ta-

Table 4.1: Comparison results of OCEAN, base LLMs (Base), and supervised fine-tuning (SFT), on three **Multi-hop Question-answering** tasks. We report with context (**w/ ctx**) and without context (**w/o ctx**) answer results with the Exact Match (EM) metric on **HotpotQA** and the Accuracy metric on **StrategyQA**. Performance on **MuSiQue** dataset is EM with context. We also use each test/validation split for each dataset and report policy evaluation $\hat{V}(\theta)$ results. We highlight the best-performed metric in **bold font** and the second-best underline for each task.

Model	Method	HotpotQA			MuSiQue		StrategyQA		
		w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$
Llama-3	Base	32.78	33.54	-10.35	11.59	-9.90	77.73	59.53	-9.25
	SFT	8.22 (-24.56)	16.49 (-17.05)	-22.28	1.80 (-9.79)	-17.09	66.52 (-11.21)	51.82 (-7.71)	-15.17
	OCEAN	33.38 (+0.6)	**33.75** (+0.21)	**-8.10**	11.67 (+0.08)	**-9.77**	75.40 (-2.33)	59.83 (+0.3)	**-5.53**
Gemma-2	Base	26.33	18.58	-31.88	5.84	-26.41	76.71	60.99	-14.06
	SFT	29.75 (+3.42)	15.91 (-2.67)	-46.92	**12.53** (+6.69)	-40.25	64.77 (-11.94)	51.97 (-9.02)	-23.27
	OCEAN	26.20 (-0.13)	19.70 (+1.12)	**-26.43**	6.87 (+1.03)	**-22.15**	74.24 (-2.47)	**66.23** (+5.24)	**-13.52**
Phi-3.5	Base	32.13	26.14	-19.49	11.85	-15.30	73.51	58.37	-13.87
	SFT	21.99 (-10.14)	7.87 (-18.27)	-44.57	6.01 (-5.84)	-42.10	63.03 (-10.48)	50.95 (-7.42)	-21.66
	OCEAN	**35.13** (+3.0)	26.23 (+0.09)	**-14.84**	10.82 (-1.03)	**-13.47**	72.20 (-1.31)	57.64 (-0.73)	**-12.25**
Mistral-0.2	Base	26.82	28.13	-19.08	5.67	-6.40	**79.33**	58.22	-11.36
	SFT	20.88 (-5.94)	14.49 (-13.64)	-18.53	7.73 (+2.06)	-12.24	52.40 (-26.93)	51.53 (-6.69)	-15.39
	OCEAN	27.24 (+0.42)	27.54 (-0.59)	**-3.12**	5.15 (-0.52)	**-5.94**	77.29 (-2.04)	56.62 (-1.6)	**-11.21**

ble 4.2. Comparing SFT and Base LLMs, we observe that directly aligning knowledge graphs with LLMs may suffer from domain change and knowledge inconsistency when downstream tasks require specific domain knowledge that could conflict with the knowledge graph in the fine-tuning stage. We also observe that SFT achieves 4.85% and 0.55% average improvements of base models on the PubMedQA dataset, with and without context respectively, whereas it suffers from 29.60%, 8.35%, 13.6% average performance decreases on the remaining tasks. Such significant discrepancies in SFT's effects on different downstream tasks further show the risk in direct knowledge editing in LLMs.

With the enhancement of OCEAN, question-answering accuracy of knowledge-intensive tasks generally improved, while OCEAN fine-tuned LLMs achieving the best performance on all three datasets, except for PubMedQA without context where SFT achieves better performance due to knowledge transfer from knowledge graph dataset. We also observe consistent policy value improvement on PubMedQA and SciQA, where the original policy values of base LLMs are relatively lower. For tasks like ARC, which does not require additional reference knowledge from context and reasoning in an easier chain of thought, OCEAN still maintains comparable policy value to the

Table 4.2: Comparison results of OCEAN, base LLMs (Base), and supervised fine-tuning (SFT), on three **Knowledge-intensive Question-answering** tasks. We report answers with context (**w/ ctx**) and without context (**w/o ctx**) on Exact Match (EM) metric on **PubMedQA** and **SciQA**. The EM performance on **ARC** dataset is without context. We also use the test/validation split for each dataset to report estimated policy values $\hat{V}(\theta)$. We highlight the best metric in **bold font** for each task.

Model	Method	ARC		PubMedQA			SciQA		
		w/o ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$	w/ ctx (%)	w/o ctx (%)	$\hat{V}(\theta)$
Llama-3	Base	79.93	**-10.38**	63.60	58.60	-25.40	83.10	57.10	-22.91
	SFT	61.87 (-18.06)	-18.42	**75.80** (+12.2)	58.00 (-0.6)	-26.03	67.10 (-16.0)	35.80 (-21.3)	-23.84
	OCEAN	80.60 (+0.67)	-12.45	66.00 (+2.4)	**59.80** (+1.2)	**-9.37**	83.20 (+0.1)	57.70 (+0.6)	**-16.63**
Gemma-2	Base	65.89	**-15.36**	34.40	40.60	-24.61	76.50	47.10	**-26.60**
	SFT	18.06 (-47.83)	-25.22	35.60 (+1.2)	21.00 (-19.6)	-26.55	79.80 (+3.3)	51.50 (+4.4)	-36.61
	OCEAN	63.21 (-2.68)	-16.20	44.60 (+10.2)	41.60 (+1.0)	**-18.72**	72.20 (-4.3)	47.50 (+0.4)	-26.77
Phi-3.5	Base	87.29	**-7.86**	70.40	41.80	-28.48	83.50	58.90	-14.46
	SFT	65.22 (-22.07)	-9.02	62.40 (-8.0)	50.20 (+8.4)	-28.40	76.90 (-6.6)	43.80 (-15.1)	-14.62
	OCEAN	**87.63** (+0.34)	-7.94	68.40 (-2.0)	47.60 (+5.8)	**-11.45**	**84.70** (+1.2)	**63.50** (+4.6)	**-13.40**
Mistral-0.2	Base	73.91	**-9.99**	51.60	36.20	-13.01	78.50	58.00	**-11.77**
	SFT	43.48 (-30.43)	-13.99	65.60 (+14.0)	50.20 (+14.0)	-21.87	64.40 (-14.1)	35.50 (-22.5)	-21.86
	OCEAN	68.90 (-5.01)	-10.89	52.60 (+1.0)	33.20 (-3.0)	**-12.42**	79.10 (+0.6)	58.40 (+0.4)	-12.00

base LLM, which demonstrates the robustness and generalizability of the proposed method.

4.5.4 Commonsense Reasoning

Finally, to demonstrate OCEAN's generalizability in preserving commonsense knowledge and preventing knowledge catastrophic forgetting (Luo et al., 2023), we evaluate OCEAN with base LLMs (Base) and supervised fine-tuning (SFT) on five commonsense reasoning tasks in Table 4.3. Since such tasks do not require chain-of-thought generation or external domain knowledge, we only evaluate the accuracy of the model's generated answers. We observe the high impact of direct SFT on knowledge graph on LLMs in potential catastrophic forgetting of commonsense knowledge, especially for the backbone LLMs of Gemma-2 and Mistral-0.2. In contrast, we show that OCEAN achieves robust performance on commonsense reasoning leverage off-policy evaluation and optimization from knowlege graph's feedback. OCEAN manages to maintain comparable performance of base LLMs (*e.g.*, Llama-3 and Phi-3.5), which have strong zero-shot commmonsense reasoning abilities. In addition, we observe that for base LLM with relatively lower performance (*e.g.*, Gemma-2 and Mistral-0.2), OCEAN enables consistent improvements. Therefore, OCEAN serves as a

Table 4.3: Comparison results of OCEAN, base LLMs (Base), and supervised fine-tuning (SFT), on five **Commonsense Reasoning** tasks. We report the Exact Match (EM) metric on these tasks and the average performance. We highlight the best method in **bold font** for each task and LLM.

Model	Method	CSQA	CSQA-2	CSQA-COT1000	OpenBookQA	Winogrande	Average
Llama-3	Base	**65.03**	**71.39**	69.50	58.80	**43.09**	**61.56**
	SFT	51.19 (-13.84)	57.06 (-14.33)	49.00 (-20.5)	63.20 (+4.4)	34.73 (-8.36)	51.04
	OCEAN	**65.03** (0.0)	68.60 (-2.79)	**72.00** (+2.5)	**60.40** (+1.6)	41.36 (-1.73)	61.48
Gemma-2	Base	57.99	62.57	63.50	51.80	49.64	57.10
	SFT	14.66 (-43.33)	65.80 (+3.23)	15.00 (-48.5)	7.40 (-44.4)	50.04 (+0.4)	30.58
	OCEAN	**67.73** (+9.74)	**63.56** (+0.99)	**72.50** (+9.0)	**57.20** (+5.4)	**50.12** (+0.48)	**62.22**
Phi-3.5	Base	68.55	**64.70**	72.50	**72.40**	**50.51**	**65.73**
	SFT	69.94 (+1.39)	61.47 (-3.23)	72.00 (-0.5)	69.40 (-3.0)	50.51 (0.0)	64.66
	OCEAN	**69.62** (+1.07)	62.77 (-1.93)	**73.50** (+1.0)	71.20 (-1.2)	50.12 (-0.39)	65.44
Mistral-0.2	Base	61.18	68.48	65.00	**64.00**	46.25	60.98
	SFT	35.87 (-25.31)	22.47 (-46.01)	33.00 (-32.0)	34.80 (-29.2)	29.12 (-17.13)	31.05
	OCEAN	**63.80** (+2.62)	**69.19** (+0.71)	**67.00** (+2.0)	62.60 (-1.4)	**46.49** (+0.24)	**61.82**

robust off-policy alignment paradigm to incorporating knowledge graph reasoning without affecting the generalizability and behaviors of pretrained LLMs.

4.6 Analysis

4.6.1 In-context Learning & Instruction tuning

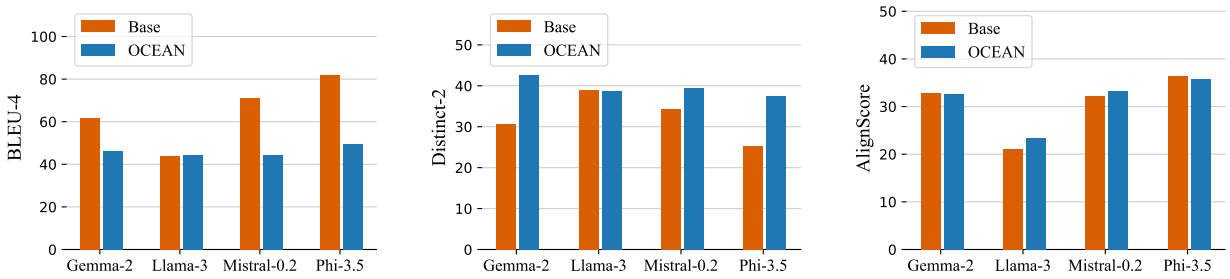
We conduct further analysis to compare the performance of both the base model and our proposed model in the scenarios of In-Context Learning and instruction fine-tuning. Specifically, we conduct this analysis using the Gemma-2 and Phi-3.5 models across three benchmark datasets: SST2 (Socher et al., 2013) for sentiment classification, AgNews (Zhang et al., 2015) for topic classification, and BoolQ (Clark et al., 2019) for reading comprehension. In the In-context Learning setup, we provide the model with a single example for each task in the prompt. For the Instruction tuning experiments, we apply LoRA (Hu et al.) to the pre-trained model and fine-tune it on each dataset for 10 epochs. Throughout these experiments, the rank parameter in LoRA is fixed at 16, and we set α in LoRA to 32 across all tasks. The results of the In-context Learning and Instruction Tuning are presented in Table 4.4. Overall, we observe that the performance of the base model and our proposed model is largely comparable across most scenarios, except in the AG News task with

Gemma-2, where OCEAN demonstrates greater performance after instruction tuning.

Table 4.4: Performance Comparison of In-Context Learning and Instruction Tuning. All datasets consist of classification tasks or true/false questions, so accuracy is used to evaluate the performance. The performance of the base model and our proposed model is largely comparable.

Model	Method	In-Context Learning				Instruction-Tuning			
		SST2	BoolQ	AG News	Avg.	SST2	BoolQ	AG News	Avg.
Gemma-2	Base	87.16	56.12	16.47	53.25	96.21	69.03	47.03	70.76
	OCEAN	89.33	55.72	13.14	52.73	96.56	68.66	60.08	75.10
Phi-3.5	Base	41.28	60.06	31.89	44.41	96.44	68.13	86.43	83.67
	OCEAN	40.48	59.11	32.37	43.98	96.44	68.81	86.24	83.83

4.6.2 Evaluation of Generation Quality Post Alignment



(a) comparison on **Self-BLEU** (b) comparison on **Distinct-2** (c) comparison on **AlignScore**

Figure 4.2: Comparison results of base LLMs and OCEAN on three evaluation metrics, Self-BLEU, Distinct-2, and AlignScore. Lower Self-BLEU scores and higher Distinct-2 scores indicate better diversity of the generated text, while higher AlignScore indicates better faithfulness in the generated answers.

To further evaluate the generation quality of post-alignment LLMs, we use the *Self-BLEU* (Zhu et al., 2018) and *Distinct-2* (Li et al., 2015) scores to evaluate the diversity of the generation, concerning the similarity between generated texts and the uniqueness of generated 2-gram phrases respectively. In addition, *AlignScore* (Zha et al., 2023) is used to evaluate the faithfulness of the generated answer given the question context. The results are presented in Figure 4.2, which show that post-alignment LLMs achieve comparable or better performances in terms of generation diversity and faithfulness. This demonstrates that while OCEAN aligns chain-of-thought reasoning with KGs, we maintain the text generation qualities of LLMs.

4.6.3 Case Study

In the previous Section 4.5.2 and 4.5.3 we observe efficient chain-of-thought alignment with improvement on the estimated policy value $\hat{V}(\theta)$. To further understand the effects of the alignment, we choose backbone LLMs, Gemma-2 and Llama-3, with significant improvements on $\hat{V}(\theta)$, and perform a sample analysis by comparing the outputs of the base model and OCEAN on the same set of questions. Our findings demonstrate that the application of our method enhances the precision and conciseness of the chain of thought in the generated responses. Some illustrative examples are provided in Figure 4.3. Specifically, in the first example, the base Llama-3 model incorrectly claims that singing is not a primary action associated with playing the guitar, which leads to an erroneous solution to the question. In contrast, our method enables the model to recognize that singing is a common activity when playing the guitar, while also understanding that making music serves as a broader term. In the second example, although both the base model and OCEAN on Gemma-2 provide reasonable answers to the question, our model demonstrates a more concise chain of thought, streamlining the reasoning process and arriving at the solution with greater simplicity.

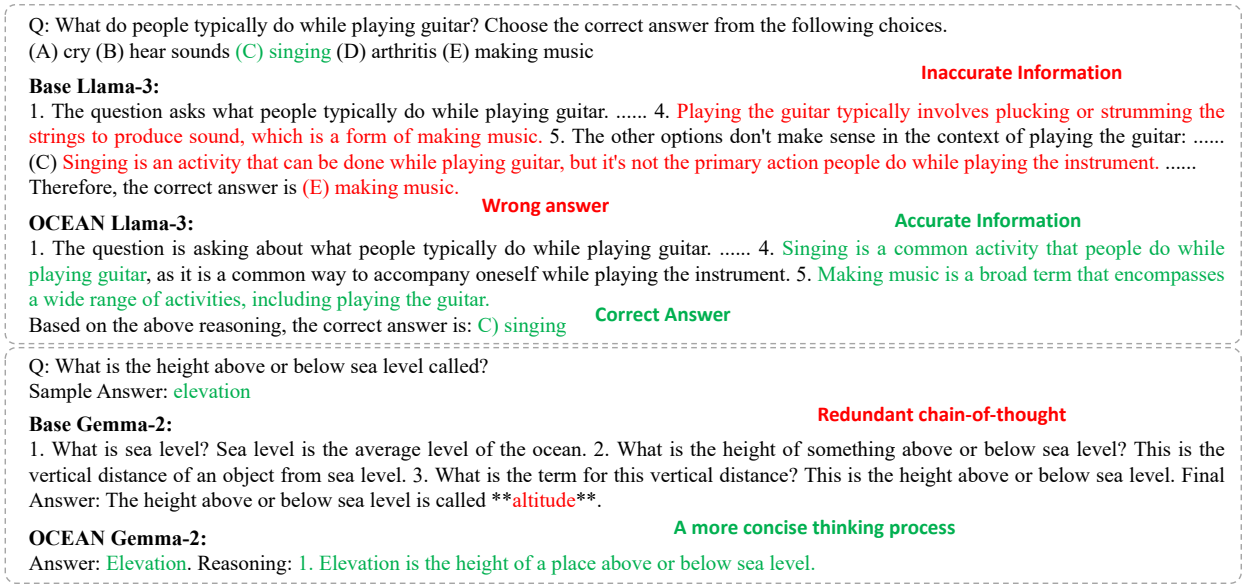


Figure 4.3: Sample comparison between the base model and OCEAN on Llama-3 and Gemma-2. Our method enables a more precise and concise Chain of thought.

4.7 Chapter Summary

In this work, we propose OCEAN to address the challenge of offline chain-of-thought evaluation and optimization of LLMs. By modeling the knowledge-graph preference and deriving feedback by developing a policy that verbalizes knowledge-graph trajectories, we propose KG-IPS estimator to estimate policy values in the alignment of reasoning paths with knowledge graphs. Theoretically, we proved the unbiasedness of the KG-IPS estimator and provided a lower bound on its variance. Empirically, our framework effectively optimizes chain-of-thought reasoning while maintaining LLMs' general downstream task performance, offering a promising solution for enhancing reasoning capabilities in large language models.

This chapter, in full, is a reprint of the material as it appears in "Offline Chain-of-Thought Evaluation and Alignment via Knowledge Graph Exploration" by Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, Jingbo Shang, and Julian McAuley, published in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 5

In-context Ranking Preference Optimization (IRPO)

5.1 Introduction

Recent advancements in Direct Preference Optimization (DPO) (Rafailov et al., 2023) allow large language models (LLMs) to compare and optimize the pairwise margin (Meng et al., 2024; Wu et al., 2024g) between positive and negative responses without explicit reward functions. However, in real-world applications (*e.g.*, conversational recommendation (Huang et al., 2025a,b; Surana et al., 2025), generative retrieval (Li et al., 2025b)), such feedback is typically collected by presenting users with an ordered in-context ranking list and asking them to select relevant items (He et al., 2023; Xie et al., 2024b) rather than providing detailed pairwise comparisons, which highlights the need for frameworks that support natural and flexible feedback formats. Such in-context feedback on ranked lists yields sparse preference signals that are not directly comparable as explicit pairwise preferences. In addition, modeling natural and flexible ranking feedback effectively requires capturing both item relevance and positional importance, of which conventional DPO methods and their underlying preference models (*e.g.*, (Rafailov et al., 2023; Chen et al., 2024a; Liu et al., 2024d)) are limited in modeling directly. Existing works enable approximations of Plackett-Luce (PL) models for ranking feedback by averaging pairwise Bradley-Terry

(BT) comparisons without directly modeling the PL distributions (Zhu et al., 2024a; Chen et al., 2024a; Liu et al., 2024d). Directly applying supervised fine-tuning (SFT) to LLMs is also insufficient for addressing this type of feedback, since ranked-list interactions inherently produce discrete and non-differentiable signals, making direct gradient-based optimization challenging.

To address these limitations of existing approaches, we propose an In-context Ranking Preference Optimization (IRPO) framework that integrates design choices across modeling and optimization to better align with real-world ranking feedback. IRPO first captures the natural form of user interactions, where users select relevant items from an in-context ranked list without providing exhaustive pairwise comparisons. To model such feedback, IRPO employs a PL-inspired positional preference model, which allows the framework to interpret sparse listwise signals by considering both item relevance and positional importance. While this formulation improves modeling fidelity, directly optimizing such objectives remains challenging due to the discrete and non-differentiable nature of common ranking metrics. To overcome this, IRPO introduces a differentiable objective based on the positional aggregation of pairwise item preferences, enabling effective gradient-based optimization.

To understand the optimization behavior of IRPO, we conduct gradient analysis and provide theoretical insights into its connection to importance sampling gradient estimation. Specifically, we show that IRPO acts as an adaptive mechanism that automatically prioritizes items with significant discrepancies between learned and reference policies, resulting in efficient and stable optimization. In addition, we derive an importance-weighted gradient estimator and show that it is unbiased with reduced variance. Empirically, we evaluate IRPO across diverse ranking tasks, including conversational recommendation, generative retrieval, and question-answering re-ranking, with various LLM backbones. Our results consistently show that IRPO significantly improves ranking performance. We summarize our contributions as follows:

- We propose In-context Ranking Preference Optimization (IRPO), a novel framework extending Direct Preference Optimization (DPO) that directly optimizes sparse and in-context ranking feedback.

- Specifically we incorporate both graded relevance and positional importance into preference optimization, addressing challenges posed by discrete ranking positions.
- We provide theoretical insights linking IRPO's optimization to gradient estimation techniques, demonstrating its computational and analytical advantages.
- Extensive empirical evaluations across diverse ranking tasks demonstrate that IRPO achieves consistent performance improvement.

5.2 Related Work

5.2.1 Ranking Generation with LLMs

Recent work has leveraged LLMs for ranking across diverse applications—including sequential (Luo et al., 2024) and conversational (Yang and Chen, 2024; Xia et al., 2023) recommendation (Li et al., 2025a), document retrieval (Liu et al., 2024b; Hu et al., 2025), and pair-wise document relevance judgments (Zhuang et al., 2023). Most approaches exploit LLMs' domain-agnostic strengths rather than improving their intrinsic ranking capabilities (Wu et al., 2024h; Pradeep et al., 2023), though fine-tuning has been recently explored (Luo et al., 2024). To our knowledge, we are the first to enhance LLM rankings through an alignment framework. Meanwhile, in contrast to prior works that focus on settings where candidate items are available (Chao et al., 2024), our framework is general and applies to cases where LLMs directly generate a list of responses from input.

5.2.2 Direct Preference Optimization for Ranking

Recent work aligns language models with human feedback via direct preference optimization (Rafailov et al., 2023; Meng et al., 2024; Wu et al., 2024g). Building on learning-to-rank methods (Valizadegan et al., 2009; Wu et al., 2024a, 2021b), several approaches recast preference alignment as a ranking task (Xie et al., 2025; Zhang et al., 2024e). For example, GDPO (Yao et al., 2024a) handles feedback diversity through group-level preferences, and S-DPO (Chen et al.,

2024a) extends these ideas to recommendation systems using multiple negatives and partial rankings. However, these approaches generally assume fully supervised or explicitly labeled feedback and require multiple forward passes for optimization, limiting their applicability to more realistic scenarios with sparse and implicit feedback.

Several recent approaches, including Ordinal Preference Optimization (OPO) (Zhao et al., 2024b), Direct Ranking Preference Optimization (DRPO) (Zhou et al., 2024b), and Listwise Preference Optimization (LiPO) (Liu et al., 2024d), explicitly leverage differentiable surrogates of ranking metrics such as NDCG to guide optimization. OPO uses a NeuralNDCG surrogate built from differentiable sorting (e.g., NeuralSort with Sinkhorn scaling) to align gains with positional discounts, thereby encoding positional importance in a smooth objective. DRPO goes further by explicitly introducing differentiable sorting networks and a margin-based Adaptive Rank Policy Score, and then trains with a differentiable NDCG objective. In contrast, LiPO frames alignment as a listwise learning-to-rank problem and commonly employs *lambda*-weighted pairwise objectives (LiPO-λ) to approximate listwise metrics like NDCG while operating over all pairs in a list.

In contrast, *IRPO* addresses a fundamentally distinct and practically motivated framework that extends the DPO (Rafailov et al., 2023) objective by jointly modeling graded relevance and positional importance, and by directly optimizing margins within a single in-context ranking list. Unlike prior works, IRPO does not rely on sorting networks or differentiable sorting approximations, and is designed to eliminate multiple forward passes, enabling a single forward pass per ranked list and lower computational cost.

5.3 Preliminaries

5.3.1 Direct Preference Alignment

DPO (Rafailov et al., 2023) enable reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022) without explicit reward modeling. Suggested by the Bradley-Terry-Luce (BTL) model of human feedback (Bradley and Terry, 1952),

a response y_1 with a reward of $r(x, y_1)$ is preferred over a response y_2 with a reward of $r(x, y_2)$ with the probability:

$$p^*(y_1 \succ y_2 \mid x) = \sigma\left(r(x, y_1) - r(x, y_2)\right), \tag{5.1}$$

where $\sigma(z) = 1/(1 + \exp[-z])$ is the sigmoid function. Suggested in (Rafailov et al., 2023), the optimal policy in the original RLHF maximization problem with a reference policy $\pi_{\text{ref}}(y \mid x)$ is given by

$$\pi_\theta(y \mid x) = \frac{1}{Z(x)}\, \pi_{\text{ref}}(y \mid x) \exp\left(\frac{1}{\beta} r(x, y)\right), Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \cdot \exp\left(\frac{1}{\beta} r(y, x)\right), \tag{5.2}$$

where the partition function $Z(x)$ serves as a normalizer (Rafailov et al., 2023). By rearranging (5.2), the implicit reward model can be derived and plugged into (5.1),

$$r(y, x) = \beta\, \log \frac{\pi^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta\, \log Z(x),$$

which formulates the maximum likelihood objective for the target policy π_θ as follows,

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_1, y_2) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_1 \mid x)}{\pi_{\text{ref}}(y_1 \mid x)} - \beta \log \frac{\pi_\theta(y_2 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}\right)\right],$$

where $Z(x)$ cancels out and thus directly optimizes the target policy without an explicit reward function.

5.3.2 Discounted Cumulative Gain

Discounted cumulative gain (DCG) has been widely used in various information retrieval and ranking tasks as a metric to capture the graded relevance of items in a ranked list while accounting for their positions (Jeunen et al., 2024; Agarwal et al., 2019). Consider a set of candidate items $\boldsymbol{e} = [e_1, e_2, \ldots, e_n]$ in a ranking list, where these items represent the potential outputs or responses for a prompt x. To rank these candidates, τ is a permutation over $\{1, 2, \ldots, n\}$ such that the item at rank i is given by $e_{\tau(i)}$. where the ranking vector is defined as $\boldsymbol{k} = [k_1, k_2, \ldots, k_n],\ k_i = \tau(i)$,

so that $e_{k_i} = e_{\tau(i)}$ is the item placed at the i-th position in the ranked list (Plackett, 1975).

For each item e_{k_i}, users may assign a relevance label y_{k_i} with explicit or implicit feedback (Järvelin and Kekäläinen, 2002), where its gain is scaled as $G(y_{k_i}) = 2^{y_{k_i}} - 1$. In addition, a positional discount factor $d(i) = 1/\log_2(1+i)$ models the decreasing importance of items placed lower in the ranking list, which leads to a weighted gain at position i given as follows,

$$w(i) = G(y_{k_i}) \cdot d(i) = \frac{2^{y_{k_i}} - 1}{\log_2(1+i)}. \tag{5.3}$$

The quality of a ranked list is measured using the Discounted Cumulative Gain (DCG):

$$\text{DCG}(\tau) = \sum_{i=1}^{n} w(i) = \sum_{i=1}^{n} \frac{2^{y_{\tau(i)}} - 1}{\log_2(1+i)}$$

where $w(i)$ represents the gain at rank i for the item with relevance $y_{\tau(i)}$.

5.4 IRPO: In-context Ranking Preference Optimization

5.4.1 Preference Modeling

To capture listwise preferences within the DPO framework, for each position i in a ranking list τ, we provide a positional preference model based on pairwise comparisons. Following the Plackett-Luce (PL) preference model (Plackett, 1975; Luce et al., 1959), for the item $e_{\tau(i)}$ at rank i,

$$p^*\Big(e_{\tau(i)} \succ \{e_j\}_{j \neq \tau(i)} \mid x\Big) = \sigma\bigg(-\log \sum_{j=1}^{n} \exp\Big(s(e_j \mid x) - s(e_{\tau(i)} \mid x)\Big)\bigg), \tag{5.4}$$

where $\sigma(z) = 1/(1+\exp[-z])$ is the sigmoid function and $s(e \mid x)$ is a score function quantifying the quality of item e given the prompt x. In (5.4), we aggregate the in-context pairwise differences between the score of $e_{\tau(i)}$ and all candidate items e_j.

To evaluate the entire ranked list, we further aggregate the individual positional preferences

into an overall list preference model by weighting each position by its NDCG gain:

$$P^*(\tau \mid x) \propto \prod_{i=1}^{n} \Big[p^*\Big(e_{\tau(i)} \succ \{e_j\}_{j \neq \tau(i)} \mid x \Big) \Big]^{w(i)}, \tag{5.5}$$

whose log-likelihood can be derived as follows,

$$\log P^*(\tau \mid x) = \sum_{i=1}^{n} w(i) \cdot \log \sigma \Bigg(- \log \sum_{j=1}^{n} \exp\Big(s(e_j \mid x) - s(e_{\tau(i)} \mid x) \Big) \Bigg), \tag{5.6}$$

which ensures that every rank contributes to the assessment of the entire ranking list.

5.4.2 Policy Optimization Objective

Following DPO (Rafailov et al., 2023) we plug the reward-based score from (5.2) into the positional NDCG preference model in (5.4), to derive our policy optimization objective for IRPO that incorporates graded relevance and the positional importance. Taking the expectation over the data distribution$(x, \boldsymbol{y}) \sim D$ and summing over all ranks, the final IRPO objective is given by:

$$\mathcal{L}_{\text{IRPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,\boldsymbol{y})} \left[\sum_{i=1}^{n} w(i) \cdot \log \sigma(z_i) \right], \quad w(i) = \frac{2^{y_{\tau(i)}} - 1}{\log_2(1+i)}, \tag{5.7}$$

where $w(i)$ is the NDCG gain at rank i according to (5.3), while the individual rank's preference z_i is

$$z_i = -\log \sum_{j=1}^{n} \exp\Big(\beta \Big[\log \frac{\pi_\theta(e_j \mid x)}{\pi_{\text{ref}}(e_j \mid x)} - \log \frac{\pi_\theta(e_{\tau(i)} \mid x)}{\pi_{\text{ref}}(e_{\tau(i)} \mid x)} \Big] \Big). \tag{5.8}$$

Our formulation naturally extends to other ranking metrics, including P@K, MAP, MRR, and eDCG, by considering their corresponding positional importance, as a generalized framework for in-context ranking preference optimization.

5.4.3 Gradient Analysis and Theoretical Insights

We further analyze the gradient by optimizing over IRPO objective to the model parameters θ:

$$\nabla_\theta \mathcal{L}_{\text{IRPO}}(\pi_\theta; \pi_{\text{ref}}) = \beta\, \mathbb{E}_{(x,\boldsymbol{y})} \left[\sum_{i=1}^{n} w(i)(1-\sigma(z_i)) \cdot \sum_{j=1}^{n} \rho_{ij} \cdot \nabla_\theta \log \frac{\pi_\theta(e_j \mid x)}{\pi_\theta(e_{\tau(i)} \mid x)} \right], \tag{5.9}$$

with the importance weights defined as

$$\rho_{ij} = \frac{\exp\Big(\beta \Big[\log \frac{\pi_\theta(e_j|x)}{\pi_{\text{ref}}(e_j|x)} - \log \frac{\pi_\theta(e_{\tau(i)}|x)}{\pi_{\text{ref}}(e_{\tau(i)}|x)}\Big]\Big)}{\sum_{k=1}^{n} \exp\Big(\beta \Big[\log \frac{\pi_\theta(e_k|x)}{\pi_{\text{ref}}(e_k|x)} - \log \frac{\pi_\theta(e_{\tau(i)}|x)}{\pi_{\text{ref}}(e_{\tau(i)}|x)}\Big]\Big)}. \tag{5.10}$$

Following previous DPO methods (Rafailov et al., 2023; Chen et al., 2024a), the gradient term $\nabla_\theta[\log \pi_\theta(e_j \mid x) - \log \pi_\theta(e_{\tau(i)} \mid x)]$ in (5.9) optimizes the model to further distinguish between the item at rank i and the remaining items in the ranking list. Intuitively, since $\sum_j \rho_{ij} = 1$ for each position i, the weights ρ_{ij} act automatically as an adaptive mechanism that assigns higher importance to items where the discrepancy between the model and the reference policy is larger. In practice, higher importance weights prioritize the gradient contribution from items ranked most wrongly relative to the target ranking.

Inspired by the importance weights ρ_{ij}, we further link our gradient analysis to the potential gradient estimator,

$$\hat{g}(e_j) = \nabla_\theta \log \frac{\pi_\theta(e_j \mid x)}{\pi_\theta(e_{\tau(i)} \mid x)}, \tag{5.11}$$

where the random variable e_j is subordinate to the distribution of importance weights $p(e_j) = \rho_{ij}$ at rank i. In practice, the gradient calculation could be approximated through importance sampling with the distribution of $p_{i,j}$ at rank i. We further show the mean and variance properties of the gradient estimator $\hat{g}(e_j)$, which serves as an unbiased estimator of the original gradient term and is more efficient in optimization.

Lemma 5.1 (Mean Analysis)**.** *The proposed gradient estimator $\hat{g}(e_j)$ is an unbiased estimation of*

the gradient term,

$$g = \sum_{j=1}^{n} \rho_{ij} \cdot \nabla_\theta \log \frac{\pi_\theta(e_j \mid x)}{\pi_\theta(e_{\tau(i)} \mid x)}, \tag{5.12}$$

in (5.9)*, which can be achieved by importance sampling with* $p(e_j) = \rho_{ij}$ *at rank* i*.*

Lemma 5.2 (Variance Analysis)**.** *We show that the expected absolute deviation of the proposed estimator is upper bounded as*

$$\mathbb{E}[|\hat{g}(e_j) - g|] \leq \sqrt{\frac{1}{n}\|\rho_{i,j}\|_\infty \cdot \mathbb{E}\left[\|\nabla_\theta \log \frac{\pi_\theta(e_j \mid x)}{\pi_\theta(e_{\tau(i)} \mid x)}\|^2\right]} \leq L\sqrt{\frac{\|\rho_{i,j}\|_\infty}{n}},$$

where $L = \max_j \left[\nabla_\theta \log \pi_\theta(e_j \mid x) - \nabla_\theta \log \pi_\theta(e_{\tau(i)} \mid x)\right]$ *and practically clipped to maintain numerical stability (Ouyang et al., 2022; Schulman et al., 2017).*

5.5 Experiments

5.5.1 Experimental Setup

Tasks To evaluate the effectiveness of IRPO on enhancing LLMs' ranking capabilities, we adopt three tasks: conversational recommendation on Inspired (Hayati et al., 2020) and Redial (Li et al., 2018) dataset; generative (supporting evidence) retrieval on HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022b) dataset; and question-answering as re-ranking on ARC (Clark et al., 2018) and CommonsenseQA (Talmor et al., 2019) dataset. Each of the tasks requires LLM to generate a ranked list among a set of candidate answers. We provide additional detail for task sections in Sections 5.5.2 to 5.5.4.

Baselines We compare IRPO against several alignment baselines: Supervised fine-tuning (SFT), which directly optimizes model outputs from explicit human annotations without preference modeling; DPO (Rafailov et al., 2023), which optimizes models from pairwise human preferences by maximizing margins between preferred and non-preferred responses; and S-DPO (Chen et al., 2024a), an extension of DPO tailored for ranking tasks, which leverages multiple negative sam-

Table 5.1: Performance on Redial and Inspired of Conversational Recommendation, evaluated using NDCG and Recall for the top 1, 5, and 10 predictions out of 20 candidate items.

Model	Method	Redial						Inspired					
		NDCG			Recall			NDCG			Recall		
		@1	@5	@10	@1	@5	@10	@1	@5	@10	@1	@5	@10
Llama3	Base	32.5	38.2	46.3	13.0	42.6	61.2	33.8	40.8	48.7	19.8	47.7	68.1
	SFT	21.6	26.5	34.9	8.1	31.2	50.4	13.8	24.7	36.2	8.1	31.7	63.3
	DPO	37.6	42.5	50.1	14.9	46.5	64.0	37.5	44.0	51.2	21.9	50.2	69.2
	SDPO	42.0	46.1	53.5	17.3	49.6	66.6	**53.1**	**57.6**	**62.3**	**30.5**	63.3	75.5
	IRPO	**74.8**	**73.6**	**79.3**	**27.9**	**78.1**	**91.3**	45.3	55.2	**62.3**	22.4	**68.8**	**87.2**
Phi3	Base	21.1	27.4	36.8	8.7	32.1	54.0	21.1	27.4	36.8	8.7	32.1	54.0
	SFT	15.9	22.2	32.7	5.8	27.2	51.6	22.1	27.0	36.6	8.5	31.2	53.4
	DPO	22.5	28.9	38.2	9.2	33.5	55.1	27.5	34.7	42.1	14.8	41.0	60.2
	SDPO	21.9	23.2	28.2	8.9	23.8	32.9	32.5	36.6	44.8	**19.6**	41.2	63.3
	IRPO	**35.0**	**39.3**	**51.1**	**12.7**	**45.6**	**72.4**	**42.2**	**46.3**	**54.6**	17.2	**54.4**	**76.0**
Gemma2	Base	19.2	26.8	38.0	7.2	32.8	58.5	15.3	22.4	31.7	7.7	28.5	52.0
	SFT	25.6	30.2	39.3	9.6	34.4	55.6	15.0	25.9	35.5	7.5	32.9	57.9
	DPO	30.1	36.0	44.3	12.1	40.4	59.4	27.5	34.7	42.1	14.8	41.0	60.2
	SDPO	48.5	52.7	59.2	19.6	55.8	70.9	40.0	44.0	48.4	**22.7**	48.2	60.0
	IRPO	**68.8**	**71.4**	**77.3**	**25.1**	**77.0**	**90.5**	**42.2**	**46.3**	**54.6**	17.2	**54.4**	**76.0**

ples, inspired by the Plackett-Luce preference model to capture richer ranking signals. We include more implementation details in the original paper.

5.5.2 Conversational Recommendation

For conversational recommendation, we use two widely adopted datasets, Inspired (Hayati et al., 2020) and Redial (Li et al., 2018). Following (He et al., 2023; Xie et al., 2024b; Carraro and Bridge, 2024; Jiang et al., 2024a), LLMs generate ranked lists of 20 candidate movies per dialogue context. To evaluate IRPO and baselines in generating ranked conversational recommendation lists, we follow (He et al., 2023) and construct candidate movie sets for each dialogue context. These candidate movies are then assigned relevance scores reflecting their importance in calculating the NDCG gain for IRPO in (5.7), where ground-truth movies receive a score of 2, GPT-generated movies a score of 1, and random movies a score of 0. Following (He et al., 2023; Xie et al., 2024b; Carraro and Bridge, 2024; Jiang et al., 2024a), we report the Recall and NDCG at top-k positions.

Table 5.2: Performance on HotpotQA and MuSiQue for Generative Retrieval, evaluated using NDCG and Recall for the top 1, 3, and 5 predictions out of 10 candidate contexts.

Model	Method	HotpotQA						MuSiQue					
		NDCG			Recall			NDCG			Recall		
		@1	@3	@5	@1	@3	@5	@1	@3	@5	@1	@3	@5
Llama3	Base	21.8	31.3	38.2	19.1	38.4	54.4	42.5	43.5	51.9	18.8	44.9	60.7
	SFT	35.3	41.4	48.3	29.5	46.3	62.1	36.9	39.1	48.0	16.6	40.7	57.6
	DPO	31.2	39.5	46.3	27.3	45.6	61.0	51.1	53.9	60.8	21.6	55.8	68.9
	SDPO	41.3	48.9	54.6	36.2	54.7	68.1	58.6	**60.8**	66.7	26.0	**62.3**	73.3
	IRPO	**94.6**	**97.2**	**97.5**	**83.6**	**99.1**	**99.8**	**65.1**	59.4	**69.4**	**27.6**	59.2	**78.0**
Phi3	Base	24.0	32.1	39.4	21.2	38.3	55.8	23.9	24.8	31.8	9.8	26.0	39.2
	SFT	26.0	34.1	41.1	23.2	40.2	56.3	33.0	32.2	41.3	14.1	33.4	50.5
	DPO	45.8	52.5	57.6	40.6	57.7	69.5	31.6	34.4	43.5	13.5	36.5	53.7
	SDPO	45.5	52.3	57.4	40.2	57.4	69.3	**41.9**	**43.9**	52.0	18.4	45.5	60.7
	IRPO	**61.2**	**73.8**	**78.5**	**53.8**	**82.9**	**93.7**	41.5	**43.9**	**52.4**	**18.5**	**45.7**	**61.7**
Gemma2	Base	32.5	39.6	45.4	28.5	45.1	58.4	40.9	42.8	51.1	17.9	44.3	59.9
	SFT	19.7	27.8	35.4	16.6	33.7	51.1	39.8	43.3	51.7	17.2	45.3	61.3
	DPO	69.1	72.6	75.5	61.5	75.3	81.9	40.9	43.1	51.1	18.1	44.9	59.8
	SDPO	53.3	59.0	63.2	47.7	62.9	72.5	40.6	42.3	50.6	17.9	43.7	59.4
	IRPO	**94.5**	**96.8**	**97.4**	**83.5**	**98.5**	**99.8**	**55.4**	**51.0**	**61.1**	**23.8**	**51.6**	**70.6**

Results in Table 5.1 show that supervised fine-tuning (SFT) can be harmful for most metrics and datasets compared with base model performance, due to strong popularity bias in conversational recommendation datasets, which is likely to cause model overfitting (Gao et al., 2025; Lin et al., 2022; Klimashevskaia et al., 2024). With multi-negative policy optimization (SDPO), such biasing effects could be alleviated, leading to relatively better performance compared to DPO and SFT. In addition, we show that IRPO achieves consistently better or comparable performance across datasets in conversational recommendation compared with baselines, by further considering positional importance for each item, weighted by NDCG weights based on the relevancy score feedback from users. With the complete ranking list optimized by pairwise comparative margins measured by DPO (Rafailov et al., 2023), IRPO acts automatically as an adaptive mechanism that assigns higher importance to items where the discrepancy between the model and the reference policy is larger.

5.5.3 Generative Retrieval

For generative retrieval, we evaluate IRPO using multi-hop question-answering datasets, HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022b). Following prior work (Shen et al., 2024; Xia et al., 2025c), we prompt LLMs to rank a set of candidate context paragraphs per question. We assign binary relevance scores, setting a score of 1 for supporting contexts and 0 for distractors.

Based on the comparative results in Table 5.2, SFT consistently reduces retrieval effectiveness relative to base models. This occurs because SFT tends to over-optimize policies, limiting their ability to generalize effectively to the nuanced retrieval challenges inherent in multi-hop queries. While SDPO, leveraging multi-negative sampling, occasionally achieves better top-1 performance compared with IRPO, IRPO attains substantial improvements in NDCG and Recall across the entire ranking list, demonstrating its effectiveness in explicitly modeling positional importance. By optimizing pairwise comparative margins comprehensively across entire ranking lists, IRPO adaptively prioritizes contexts with larger divergences between model predictions and preference feedback. Thus, IRPO offers robust generalizability and better retrieval accuracy for complex generative retrieval tasks.

5.5.4 Question-answering as Re-ranking

In the question-answering as re-ranking scenario, we assess the ability of LLMs to identify correct answers from multiple choices based on contextual relevance. We evaluate IRPO on two widely-used datasets, ARC (Clark et al., 2018) and CommonsenseQA (Talmor et al., 2019) . Each candidate is explicitly assigned binary relevance: the correct answer receives a score of 1, and incorrect answers a score of 0.

Comparative results summarized in Table 5.3 highlight IRPO's robust improvements across models and datasets. While SFT achieves relatively better performance compared to its performance on other tasks, it still struggles to prioritize and disambiguate the correct answer among

Table 5.3: Performance on the ARC and CommonsenseQA QA datasets, evaluated using NDCG and Recall for the top 1, 3, and 5 predictions out of 10 answer choices.

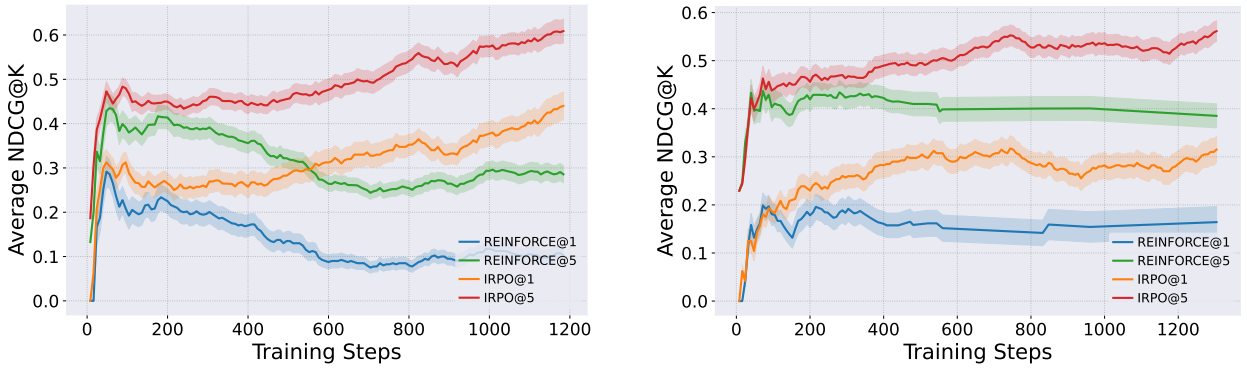
Model	Method	ARC						CommonsenseQA					
		NDCG			Recall			NDCG			Recall		
		@1	@3	@5	@1	@3	@5	@1	@3	@5	@1	@3	@5
Llama3	Base	13.2	21.7	28.6	13.2	28.4	44.9	24.1	32.6	38.9	24.1	39.3	54.8
	SFT	13.5	21.4	28.4	13.5	27.7	44.6	22.8	31.6	38.7	22.8	38.6	56.1
	DPO	15.2	22.8	30.4	15.2	29.1	47.3	54.4	59.5	63.4	54.4	63.6	73.2
	SDPO	23.4	28.3	35.4	23.4	32.5	50.0	54.2	59.9	63.9	54.2	64.4	74.2
	IRPO	**27.4**	**38.9**	**46.5**	**27.4**	**47.6**	**66.3**	**68.6**	**83.1**	**85.0**	**68.6**	**92.8**	**97.5**
Phi3	Base	9.8	17.9	25.1	9.8	24.3	41.9	24.3	33.2	38.7	24.3	40.0	53.6
	SFT	8.8	17.7	25.7	8.8	24.7	**44.3**	24.4	32.6	39.6	24.4	39.0	56.3
	DPO	9.5	17.9	25.5	9.5	24.7	43.2	**54.0**	59.7	63.5	**54.0**	64.0	73.4
	SDPO	9.7	18.0	25.2	9.7	24.7	42.4	48.1	54.3	59.0	48.1	59.0	70.4
	IRPO	**10.4**	**19.5**	**26.8**	**10.4**	**25.7**	43.8	53.1	**70.1**	**74.1**	53.1	**82.2**	**92.0**
Gemma2	Base	27.4	35.6	42.1	27.4	42.6	58.4	27.2	35.4	41.4	27.2	42.0	56.7
	SFT	23.6	32.9	39.3	23.6	40.2	55.7	29.5	38.0	44.1	29.5	44.7	59.7
	DPO	27.0	34.4	40.7	27.0	40.5	56.1	29.4	37.9	43.7	29.4	44.7	59.1
	SDPO	**28.7**	36.5	42.6	**28.7**	43.0	58.0	57.0	62.4	66.2	57.0	66.5	75.8
	IRPO	27.0	**45.5**	**53.3**	27.0	**59.8**	**78.7**	**66.3**	**79.2**	**81.7**	**66.3**	**88.3**	**94.5**

similar candidate answers. On the other hand, SDPO, benefiting from multi-negative optimization, typically yields better performance by distinguishing subtle semantic differences among distractors. IRPO further demonstrates superior effectiveness in this challenging ranking task, substantially outperforming all baseline approaches. Aligned with our theoretical insights into IRPO optimization in Section 5.4.3, IRPO boosts performance by adaptively focusing on candidates with greater discrepancies, enabling comprehensive comparisons that effectively resolve subtle semantic differences and yield substantial gains on challenging datasets like ARC and CommonsenseQA.

5.6 Analysis

In this section, we analyze the optimization behavior and performance of IRPO in both online and offline settings. For on-policy optimization (in Section 5.6.1), we extend IRPO to its online variant **Iterative IRPO**. Additionally, we present offline learning curves for IRPO across multiple

tasks and various backbone LLMs (Section 5.6.2) in Figure 5.2. We then conduct an ablation study to investigate the role of relevance and positional weighting in IRPO's objective (Section 5.6.3), demonstrating the necessity of jointly modeling these factors for effective ranking alignment. Finally, we compare IRPO with recent learning-to-rank approaches, focusing on LiPO (Liu et al., 2024d) (Section 5.6.4).



(a) ARC (b) CommonsenseQA

Figure 5.1: Comparison of REINFORCE and IRPO on ARC and CommonsenseQA.

5.6.1 On-policy Online Optimization of Iterative IRPO

We explore an on-policy variant of IRPO, **Iterative IRPO**, which adapts to an online optimization setting Hu (2025); Wu et al.; Shao et al. (2024). In this setting, models sample their responses based on queries, rather than relying on predefined ranking lists from the original datasets. These on-policy sampled responses are compared with ground-truth annotations, simulating a realistic human feedback loop. We conduct such on-policy online learning experiments on ARC and CommonsenseQA, compared to a standard policy-gradient baseline, REINFORCE (Sutton et al., 1999). In Figure 5.1, we show that the Iterative IRPO achieves constantly increasing NDCG scores, while REINFORCE fails to explore effective candidates, leading to insufficient feedback. Aligned with the design of IRPO, which prioritizes more relevant items while considering positional importance, Iterative IRPO could improve the general quality of the entire ranking list, which significantly benefits on-policy exploration performance. Supported by our theoretical insights (Section 5.4.3), we link IRPO's optimization to an efficient importance sampling method, which inherently serves

as an effective exploration mechanism when Iterative IRPO is enabled online. To further illustrate, we provide a qualitative comparison of outputs generated by the base model, Iterative IRPO, and REINFORCE in the original paper.

5.6.2 Optimization Analysis of IRPO

We further evaluate IRPO's offline optimization performance in the experimental results in Figure 5.2, showing that IRPO consistently achieves higher evaluation NDCG scores and exhibits stable optimization. across six diverse benchmarks (*Inspired*, *HotpotQA*, *ARC*, *Redial*, *MuSiQue*, and *CommonsenseQA*) using three different LLM backbones: Llama3, Gemma2, and Phi3. This robust performance improvement is attributable to IRPO's adaptive importance weighting mechanism, which effectively prioritizes gradient updates toward ranking positions with higher relevance discrepancies. This mechanism allows IRPO to rapidly capture essential ranking signals from the feedback, leading to stable and consistent optimization behavior, a finding further supported by our theoretical analysis in Section 5.4.3.

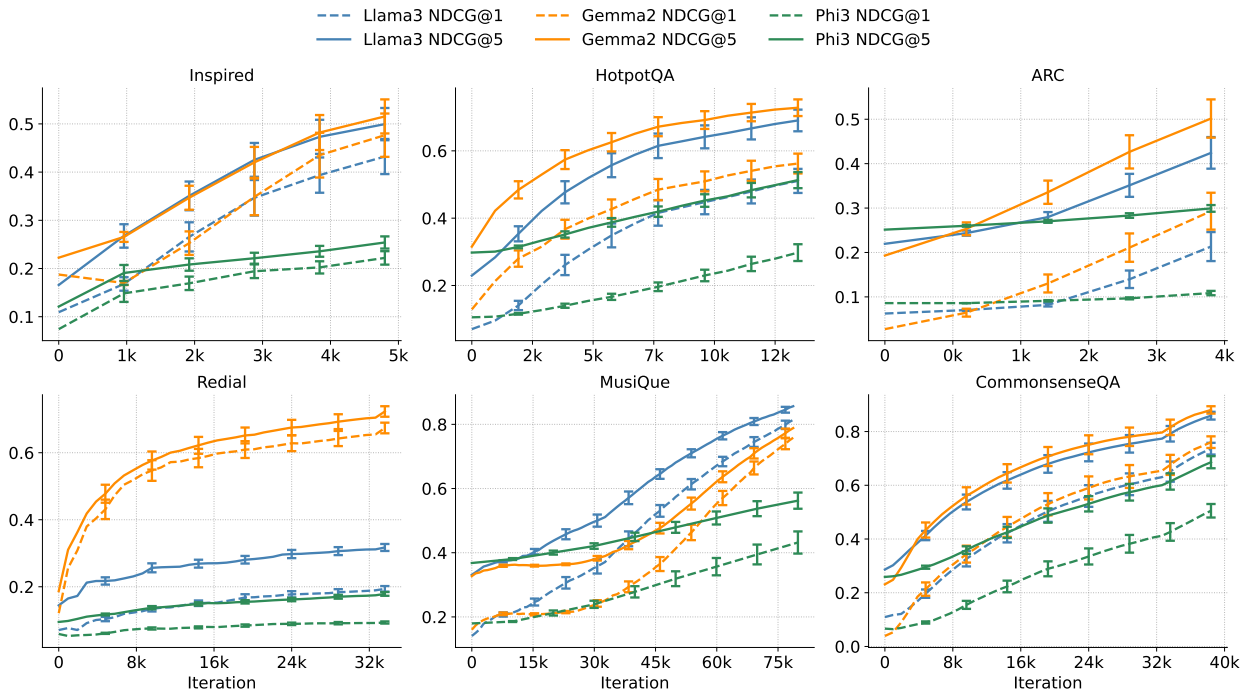


Figure 5.2: IRPO's performance across six benchmarks using Llama3, Gemma2, and Phi3

To highlight the strengths of IRPO over baseline methods, we include three representative qualitative examples in the original paper. These showcase how IRPO more effectively ranks contextually relevant and coherent responses.

5.6.3 Ablation Study

IRPO inherently prioritizes relative comparisons over absolute weight magnitudes, making it less sensitive to specific weighting schemes. To validate this, we conduct an additional ablation study evaluating two alternative positional weighting methods: (1) **abl1:** $w(i) = \frac{1}{\log(1+i)}$ (positional weights without relevance) (2) **abl2:** $w(i) = \frac{2^{y_i}-1}{i}$ (alternative relevance scaling).

As shown in the original paper, removing the relevance component (as in *abl1*) leads to significant degradation in ranking performance across all datasets. These results underscore the importance of modeling both item relevance and positional importance within IRPO's objective.

5.6.4 Comparison with Learning-to-Rank (LiPO)

Although IRPO and LiPO (Liu et al., 2024d) address distinct feedback settings, we provide a comparative analysis here to clearly contextualize IRPO within recent learning-to-rank (LTR) paradigms. LiPO (Liu et al., 2024d) is a recent representative baseline explicitly extending direct preference optimization (DPO) by integrating general learning-to-rank principles. Unlike IRPO, LiPO is designed primarily for scenarios with fully supervised or extensively labeled listwise data, allowing straightforward integration into standard supervised ranking setups.

To meaningfully compare these distinct approaches, we adapt LiPO to align closely with our sparse, in-context feedback scenario and conduct comparative experiments on representative benchmarks (ARC and MuSiQue). Results summarized in the original paper indicate that IRPO consistently outperforms LiPO, underscoring the advantage of IRPO's explicitly modeled positional relevance and sparse feedback setting.

5.7 Chapter Summary

In this work, we introduced IRPO, a novel alignment framework that directly optimizes LLMs for ranking tasks using sparse, in-context user feedback. By explicitly modeling both item relevance and positional importance within a differentiable ranking objective, IRPO effectively addresses the limitations of existing DPO methods. Our theoretical insights demonstrated IRPO's adaptive prioritization mechanism and established its connection to importance sampling, providing unbiased gradient estimation with reduced variance. Extensive empirical evaluations across conversational recommendation, generative retrieval, and question-answering re-ranking tasks consistently showed IRPO's superior ranking performance. Our findings highlight IRPO as an effective and efficient method for aligning LLMs with realistic user preferences, paving the way for broader integration into in-context action-space exploration and reinforcement learning in dynamic online settings.

This chapter, in full, is a reprint of the material as it appears in "In-context Ranking Preference Optimization" by Junda Wu, Rohan Surana, Zhouhang Xie, Yiran Shen, Yu Xia, Tong Yu, Ryan A. Rossi, Prithviraj Ammanabrolu, and Julian McAuley, published in *Conference on Language Modeling (COLM)*, 2025. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Debiasing Chain-of-Thought via Causal Intervention (DeCoT)

6.1 Introduction

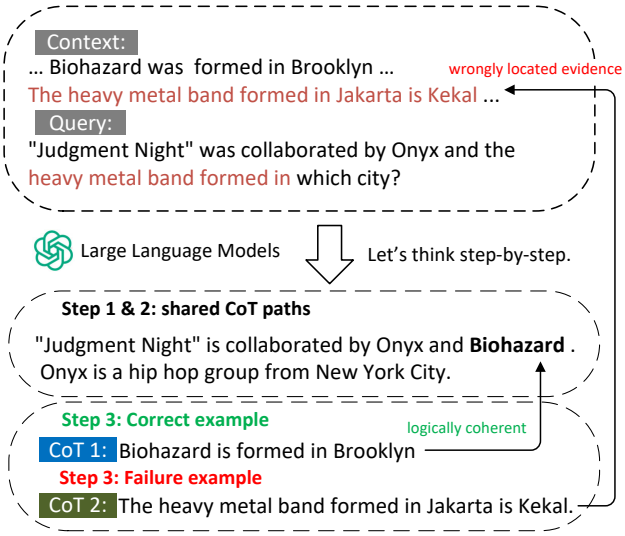


Figure 6.1: LLMs' internal knowledge bias can trigger the usage of irrelevant information in prompts, generate incoherent reasoning chains, and impair the model's logical reasoning ability. This example is derived from the experiments by GPT3.5 on HotpotQA where the context is the one provided in the dataset. Please note that 'The heavy metal band formed in Jakarta is Kekal' refers to a heavy metal band that is different from Biohazard. However, GPT3.5 incorrectly assumes that 'The heavy metal band'.refers to Kekal, and provides incorrect information in step 3 of the CoT.

For knowledge-intensive tasks Petroni et al. (2021); Hu et al. (2023); Sun et al. (2023b), specific knowledge is required to obtain an accurate response, which can be out of the distribution

of LLMs' internal knowledge Yao et al. (2023a); Yuan et al. (2023d). Since frequently fine-tuning LLMs can be highly expensive and inefficient Zhai et al. (2023), the LLM's internal knowledge can also be outdated and cause knowledge bias problems in LLMs Zhang et al. (2023d); Wu et al. (2023b); Zhang et al. (2023e). To efficiently incorporate external knowledge (*i.e.*, context), methods are proposed to retrieve task-relevant language evidence Liu et al. (2023b); He et al. (2022); Zhu et al. (2023b); Shao et al. (2023); Trivedi et al. (2022a). Additionally, external knowledge bases can also directly augment and edit the knowledge-injected prompts Wen et al. (2023); Sun et al. (2023a); Baek et al. (2023); Zhao et al. (2023b). However, simply injecting external knowledge in prompts does not guarantee that LLMs can identify and use relevant information in the prompts Shi et al. (2023a); Weston and Sukhbaatar (2023), especially when the LLM learns biased information in pretraining data Zhang et al. (2023d). The knowledge bias in LLMs can further cause knowledge conflict or misunderstanding between external knowledge and the model's internal knowledge Mallen et al. (2023); Wang et al. (2023f,a). In such cases, LLMs may use irrelevant information from prompts and generate incorrect and unexpected responses Li et al. (2023e); Xie et al. (2023).

When the LLM relies on chain-of-thought (CoT) reasoning for complex tasks, the biased information from pretrained models further impairs LLMs' reasoning abilities. Many works propose to verify and post-edit the generated reasoning paths before prompt again Zhao et al. (2023b); Peng et al. (2023); Wang et al. (2024b) to eliminate the factual errors in the generated CoT paths. However, logical reasoning errors can not be easily detected or corrected, as the effectiveness of factual verification and post-editing reasoning chains can be limited to simply injecting more knowledge. For example in Figure 6.1, given the query (*e.g.*, ""Judgment Night" was collaborated by Onyx and the heavy metal band formed in which city?") and context which provides external knowledge Lewis et al. (2020), the LLM may generate logically incorrect CoT (*e.g.*, CoT 2), in which the last chain deviates from the reasoning paths (*e.g.*, instead of the origin of "Biohazard", some arbitrary band mentioned in the context). Such logical incoherence can be caused by the spurious correlation between the query (*e.g.*, the concept "the heavy metal band formed in") and the LLM's

internal knowledge understanding. Thus, the spurious correlation can lead the LLM to find some arbitrary evidence in the context regardless of its logical connection to the previous chain, as long as it contains the exact phrase. In such cases, factual verification methods cannot detect logical reasoning errors, and the answer can still be incorrect even with the facts verified as correct.

In this work, we propose a novel causal view via a Structural Causal Model (SCM) Pearl et al. (2016) to formally explain the internal knowledge bias of LLMs. To measure spurious correlation, we propose to use external knowledge as an instrumental variable Morgan and Winship (2015) to estimate the Average Causal Effect (ACE) of CoT reasoning paths in LLMs through causal intervention. Based on the measurement of ACE, we can further introduce a CoT sampling method to find the best CoT as a mediator and conduct front-door adjustment Pearl (2009). In this approach, the spurious correlation between LLMs' internal knowledge and task queries can be reduced, which ensures correct CoT reasoning and LLM generation. We summarize our contributions as follows:

- We discover that the bias from LLMs can trigger the usage of irrelevant information in the prompts, and cause the LLM to generate incoherent reasoning chains that impair the model's reasoning ability.
- To formally understand the bias affecting CoT reasoning abilities, we propose a novel causal view introducing the external knowledge in prompts as an instrumental variable. This causal view uncovers the spurious correlation between queries and LLMs' internal knowledge understanding.
- To alleviate the bias and ensure correct CoT reasoning, we estimate the average causal effect (ACE) between the CoT and the answer, and further propose a CoT sampling method to conduct the front-door adjustment.
- We conduct multiple experiments on various knowledge-intensive tasks as well as numbers of LLM backbone models, which demonstrates the effectiveness of our method.

6.2 Related Work

LLMs in Knowledge-intensive Tasks. In knowledge-intensive tasks Petroni et al. (2021); Hu et al. (2023); Yang et al. (2018); Welbl et al. (2018) the LLM is asked to respond based on the provided context and its intrinsic knowledge. Retrieval-augmented prompting methods focus on identifying accurate and comprehensive evidence from support documents Hoshi et al. (2023); Qian et al. (2023), in-context examples Press et al. (2022); Khattab et al. (2022); Wang et al. (2025a), knowledge bases Trivedi et al. (2022a); Xu et al. (2023); Wang et al. (2024b); Feng et al. (2023); Zhu et al. (2023a), knowledge graphs Wen et al. (2023); Sun et al. (2023a); Salnikov et al. (2023); Zhang et al. (2023b) and human feedback Zhang et al. (2023d). However, extensive knowledge-injected prompts can introduce irrelevant information to distract LLMs Shi et al. (2023a); Wang et al. (2023g) and cause LLMs' unpredictable behaviors Li et al. (2023e); Chen et al. (2023c). Instead of focusing on how to identify the best knowledge evidence, we investigate how to find logically correct CoT reasoning paths.

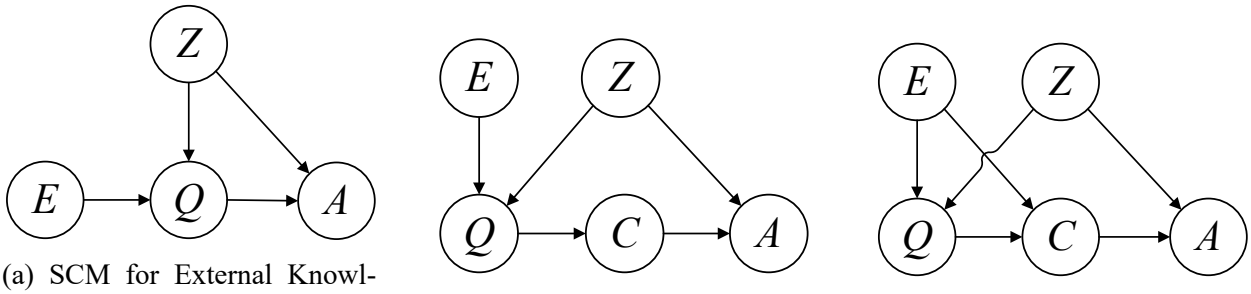
Chain-of-thought Prompting. Chain-of-thought prompting has shown great potential in explaining LLMs' thinking process Yuan et al. (2023a); Li and Du (2023) and answering multi-hop questions Wang et al. (2023c); Ma et al. (2022) in complex reasoning tasks Fu et al. (2023). However, further works also mention issues of faithfulness and self-consistency in LLMs Lanham et al. (2023); Turpin et al. (2023a). To improve the faithfulness of intermediate chains, several works propose to verify and edit He et al. (2022); Wang et al. (2023d); Zhao et al. (2023b) the factual errors in unfaithful chains, and Radhakrishnan et al. (2023); Zhu et al. (2023a) propose to decompose complex questions and answer them individually. We argue that logically incorrect reasoning paths can lead the LLM astray from the right direction of finding the answer, even with the chains factually correct. Similar to previous works, our method first generates some candidate chain-of-thought reasoning paths, which makes our method use similar numbers of API calls (or inference times) per sample.

Causal Intervention in Language Models. Causal intervention methods in language models focus

on entity-level spurious correlation Wang et al. (2023b); Zeng et al. (2020); Yang et al. (2023) or sentence-level selection bias McMilin (2022). Since LLMs are black-box models Gat et al. (2023); Cheng et al. (2023), direct methods of causal-aware model reparameterization methods are limited in usage. With causal explanability extracted from the models Wu et al. (2021a); Zhao et al. (2023a), further studies introduce human-in-the-loop debiasing methods Zhang et al. (2023d); Wu et al. (2022a, 2021a); Liu et al. (2024e). However, human effort is normally more expensive, and involving humans in the loop may reduce the efficiency of the system. Instead, our method uses counterfactual context to automatically measure the causal effect and find better CoT to improve the performance of LLMs.

6.3 A Causal View

To understand the causal relationships in knowledge-intensive tasks, we introduce a Structural Causal Model (SCM) Pearl et al. (2000) and identify the internal knowledge understanding of the LLMs (Z) as the confounder. In Figure 6.2a and Figure 6.2b, we formulate two types of conventional knowledge injection methods as two SCMs respectively. With the SCMs, we explain the effectiveness of conventional knowledge injection methods and their limitations. We further present the SCM of our method, the debiasing chain-of-thought (**DeCoT**), in Figure 6.2c. The formulation of our method, **DeCoT** is illustrated in Figure 6.3 and detailed in Section 6.5.



(a) SCM for External Knowledge (b) SCM for Chain-of-thought (c) SCM for DeCoT

Figure 6.2: Structural causal graphs for (a) injection of external knowledge, *i.e.*, context Lewis et al. (2020), (b) chain-of-thought prompting and (c) our approach using external knowledge as an instrumental variable (detailed in Section 6.5.2).

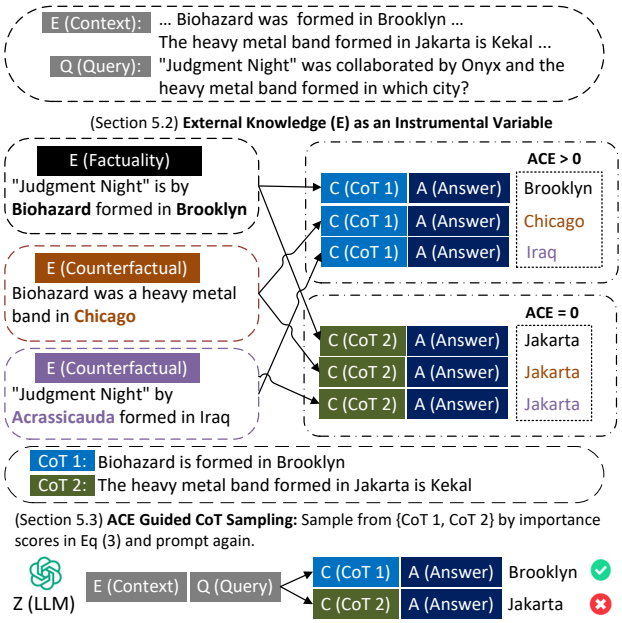


Figure 6.3: The illustration of our proposed debiasing chain-of-thought method DeCoT (detailed in Section 6.5).

6.3.1 SCM for External Knowledge

In Figure 6.2a, the causal path $E \rightarrow Q \leftarrow Z$ represents the knowledge injection process, where E denotes the external knowledge, Q denotes queries in the inference stage and Z denotes the LLM's internal knowledge. Ideally, the query (Q) alleviates spurious correlations as a collider, influenced by the external knowledge (E) and LLM's internal knowledge (Z), when E and Z are causally irrelevant Pearl et al. (2000). However, most knowledge injection methods Baek et al. (2023); Li et al. (2023d) incorporate the external knowledge as the context which is prefixed to the input prompt. Thus, the causal influence of the external knowledge on the query is also determined by the LLM, which makes E and Z dependent variables, and the spurious correlation between Q and Z remains.

6.3.2 SCM for Chain-of-thought

Chain-of-thought (as in Figure 6.2b) is introduced Li et al. (2023c); Wei et al. (2022); Fu et al. (2023) to make the LLM explain and follow the reasoning path before giving the final answer.

The causal path $Q \rightarrow C \rightarrow A$ shows that the CoT (C) can serve as the mediator between the query (Q) and the answer (A). However, since the CoT is also prompted from the LLM Wei et al. (2022); Fu et al. (2023), it can also be causally dependent on LLM's internal knowledge and thus forms the spurious correlation between C and Z. Notably, knowledge editing methods Zhao et al. (2023b); Li et al. (2023c); Peng et al. (2023) can correct the factual errors in the context and the reasoning paths, while the reasoning logic remains incorrect.

6.4 Preliminaries

6.4.1 Task Formulation

For knowledge-intensive question-answering tasks, the model is prompted with a query $Q = [q_1, q_2, \ldots, q_n]$ and a passage of context $E = [e_1, e_2, \ldots, e_l]$, *i.e.*, external knowledge Lewis et al. (2020). Given the query Q and the context E, the model θ is prompted to recurrently generate the response Y by sampling from the conditional probability distribution,

$$y_t \sim p_\theta \left(y|E, Q, y_{<t}\right).$$

Following the previous setting Welbl et al. (2017); Trivedi et al. (2022b), we make no assumptions on the context more than what is available in the context of the data samples from their datasets. In this setting, there are irrelevant contexts Tu et al. (2020); Yang et al. (2018), which lead to the spurious correlation identified in our paper. As illustrated and explained in Figure 6.2a of Section 6.3.1, the model directly generates the answer $A = [a_1, a_2, \ldots, a_m]$ without providing the intermediate reasoning process (*i.e.*, $A = Y$).

Chain-of-thought Prompting. Following Wei et al. (2022), we add the additional instruction to ask the model to generate its reasoning paths C by explaining step-by-step, before generating the final answer A (*i.e.*, $Y = [C, A]$). By sampling N different CoTs $C = [C_1, C_2, \ldots, C_N]$ conditioned on the query Q and the context E, we can further condition the generation process of the answer

A,

$$C_i \sim p_\theta\left(C|E,Q\right), \tag{6.1}$$

$$A_{i,r} \sim p_\theta\left(A|E,Q,C_i\right). \tag{6.2}$$

In Eq. (6.1), since CoTs (C) are also generated from the LLM in which the pretrained internal knowledge (Z) can also confound on the generation process. As explained in Section 6.3.2, the confounding effect can not only affect the factual accuracy of the generated CoTs but also lead to incorrect reasoning logic. Thus knowledge editing and verification methods Zhao et al. (2023b); Li et al. (2023c); Peng et al. (2023) which solve the former problem, are limited in correcting logical errors in CoTs.

6.5 DeCoT: Debiasing Chain-of-thought

6.5.1 SCM for DeCoT

In Figure 6.2c, the CoT (C) is a mediator between the query (Q) and the answer (A). Based on the front-door criterion Pearl et al. (2000), a mediator (C) should be causally independent of the confounder (Z), to enable front-door adjustment. However, in practice, CoTs are also generated by LLMs, which suggests potential spurious correlations between CoTs (C) and the LLM's internal knowledge (Z). Thus, to track the bias from the unobserved confounder Z, we introduce the external knowledge as an instrumental variable (IV) Kawakami et al. (2023); Kilbertus et al. (2020); Yuan et al. (2023c). By changing the value of the instrumental variable E (*i.e.*, the external knowledge), we estimate the true causal relationship between C and A Yuan et al. (2023c). For example, in Figure 6.3, two pieces of counterfactual external knowledge (*e.g. "Biohazard formed in Chicago"* and *"Judegment Night was by Acrassicauda formed in Iraq"*) are introduced in the same example of Figure 6.1, where the average causal effect (ACE) is calculated as in (6.4). Due to the spurious correlation in the third chain of thoughts of "CoT 2", responses generated from "CoT

2" remain unchanged (*i.e.*, ACE $= 0$), while responses generated from the correct reasoning path "CoT 1" change corresponding to counterfactual evidence (*i.e.*, ACE > 0).

6.5.2 External Knowledge as an Instrumental Variable

We model the external knowledge as an instrumental variable E to understand the causal relationship between the CoT C and the answer A Yuan et al. (2023c). Due to the limitation of directly controlling the generation process of CoTs, we perform the causal treatment by including counterfactual knowledge through the instrumental variable E. Specifically, we query the LLM to extract T factual entities $V = [v_1, v_2, \ldots, v_T]$ which correspond to T counterfactual context $E_1^*, E_2^*, \ldots, E_T^*$. In each sample

$$E_j^* = [e_1, e_2, \ldots, v_j, \ldots, e_l]\,,$$

the corresponding factual entity v_j is to be replaced by counterfactual entities. Then, the LLM is further prompted to propose P counterfactual entities $V_j^* = [v_{j,1}^*, v_{j,2}^*, \ldots, v_{j,P}^*]$ to each extracted entity $v_j \in V$. P counterfactual context samples

$$E_{j,k}^*(v_{j,k}^*) = [e_1, e_2, \ldots, v_{j,k}^*, \ldots, e_l], k \leq P \tag{6.3}$$

are constructed, by replacing the corresponding factual entity v_j in each sample E_j^*. In this approach, we estimate the average causal effect (ACE) corresponding to each CoT reasoning path C_i,

$$\begin{aligned} \text{ACE}(C_i, v_j) &= \mathbb{E}\left(A|do(E), Q, C_i\right) - \mathbb{E}\left(A|do(E_j^*), Q, C_i\right) \\ &= \mathbb{E}_{v_{j,k}^* \in V_j^*}[p_\theta\left(A|E, Q, C_i\right) - p_\theta\left(A|E_{j,k}^*(v_{j,k}^*), Q, C_i\right)], \end{aligned} \tag{6.4}$$

in which the average causal effect measures the decreased confidence in the answer (measured as in (6.2)) with counterfactual context as the evidence. We observe that the average causal effect

of different factual entities can vary to the context, queries, and CoTs, which we further conduct analysis experiments in Section 7.5.4. To consider the overall causal effect of the external context on each CoT, we propose to measure the average causal effect of all the intervened entities,

$$\text{ACE}(C_i) = \mathbb{E}_{v_j \in V} \, \text{ACE}(C_i, v_j), \tag{6.5}$$

where the intervened entities v_j are sampled from a uniform distribution of the external context E.

6.5.3 Average Causal Effect Guided Chain-of-thought Sampling

With the measured ACE scores, we develop an efficient sampling approach to obtain high-quality CoTs with more coherent reasoning chains, without additional deconfounding layers Zhang et al. (2020); Wu et al. (2022a) for LLM finetuning. Since LLMs are black-box models Gat et al. (2023); Cheng et al. (2023), direct causal intervention methods on the parameterization of the input query Q and the context E are limited. Thus, we propose to use the sampled CoTs C as the mediator variable to conduct the front-door adjustment.

Based on the measured average causal effect (ACE), we construct importance scores in terms of how the final answer A reacts to different CoTs C intervened by the context E,

$$C^* \sim \text{softmax}\left[p_\theta\left(C_i|E, Q\right) \cdot \text{ACE}(C_i)\right], \tag{6.6}$$

and the front-door adjustment can be realized by introducing the mediator C^* sampled with the largest average causal effect in the reasoning path,

$$A^* \sim P\left(A|E, Q, do(C)\right) \propto p_\theta\left(A|E, Q, C^*\right). \tag{6.7}$$

The causal effect on the sampled answer A^* is mediated by the sampled CoT reasoning path C^*, whose mediator-outcome confounding effect is controlled and alleviated. We summarize our algorithm DeCoT in the original paper.

6.6 Experiments

Table 6.1: The comparison results of DeCoT based on different backbone LLMs on four knowledge-intensive tasks. The best results for each backbone model and each dataset are highlighted in a **bold font**.

		HotpotQA		MuSiQue		SciQ		WikiHop		Average	
Model	**Decoding**	EM↑	F1↑	EM↑	F1↑	EM↑	F1↑	EM↑	F1↑	EM↑	F1↑
Flan-T5	CoT w/o ctx	7.41	17.99	2.57	8.50	11.09	17.80	4.12	6.88	6.30	12.79
	CoT	9.48	23.70	19.53	27.61	51.75	63.79	15.02	21.79	23.95	34.22
	CAD	9.65	24.77	**20.56**	28.57	59.69	69.94	17.28	24.31	26.80	36.90
	DeCoT	**11.72**	**28.70**	**20.56**	**30.54**	**63.55**	**75.64**	**22.34**	**28.41**	**29.54**	**40.82**
LlaMA-2	CoT w/o ctx	1.67	3.04	0.56	1.44	4.08	5.45	1.19	1.64	1.88	2.89
	CoT	8.86	26.79	20.22	27.46	30.64	39.59	23.10	28.23	20.71	30.52
	CAD	**10.53**	30.98	21.62	28.10	33.93	41.35	23.50	29.81	22.40	32.56
	DeCoT	10.03	**31.48**	**22.75**	**30.99**	**48.58**	**57.95**	**27.35**	**34.46**	**27.18**	**38.72**
GPT-3.5	CoT w/o ctx	5.60	30.97	2.09	7.96	29.82	43.18	11.62	19.31	12.28	25.36
	CoT	5.10	32.55	22.30	34.22	54.53	68.50	25.40	35.25	26.83	42.63
	CAD	5.43	35.24	24.14	36.81	57.74	70.73	28.37	38.45	28.92	45.31
	DeCoT	**10.21**	**40.19**	**31.28**	**44.14**	**64.61**	**78.10**	**31.89**	**43.45**	**34.50**	**51.47**

6.6.1 Dataset and Evaluation

Knowledge-intensive tasks commonly require each question to be paired with a paragraph of context as support evidence Li et al. (2023c); Zhu et al. (2023b); Su et al. (2023); Jang et al. (2023). Our method focuses on the reasoning errors caused by the spurious correlation from LLM's pretrained knowledge. We follow the setting Welbl et al. (2017, 2018); Yang et al. (2018); Trivedi et al. (2022b); Shi et al. (2023b); Wei et al. (2022) and use the original supporting contexts from the datasets without any additional data processing. We follow the evaluation protocols in Yang et al. (2018) and conduct our experiments on datasets as follows:

HotpotQA Yang et al. (2018) contains questions that require multi-step reasoning over multiple support contexts. For each question, support documents are provided in the dataset, which are used as the context in our experiments.

MuSiQue Trivedi et al. (2022b) is another multi-step question answering dataset. Similar to previ-

ous work Ramesh et al. (2023), we conduct our experiment on the challenging part of the dataset, in which questions are annotated as ≥ 4 *hops*.

WikiHop Welbl et al. (2018) is a multi-choice multi-hop reasoning dataset. We use the queries in the dataset as the questions Tu et al. (2019) in our setting. For baselines and our method, the models are prompted to generate the answers free-form instead of retrieving them from the candidate list.

SciQ Welbl et al. (2017) is a domain-specific question-answering task that contains only scientific questions. We evaluate baselines and our method on test samples with supporting evidence.

For datasets that lack test labels, we follow the same evaluation protocol as Press et al. (2022); Shao et al. (2023); Chen et al. (2023b) and use the development sets as our test set. We use the Exact Match (EM) and F1 proposed in Yang et al. (2018) as our evaluation metrics.

6.6.2 Baseline and Backbone Model

Following Shi et al. (2023b); Su et al. (2023); Trivedi et al. (2022a), we have applied our method to different pretrained LLMs: Flan-T5-XXL Chung et al. (2024) which has 11B model parameters, LlaMA-2-7B Touvron et al. (2023) and GPT-3.5 Turbo Brown et al. (2020). For LlaMA-2-7B model, we choose the finetuned versions from human feedback Christiano et al. (2017), which can generally yield more stable chain-of-thought reasoning paths.

For baselines, we compare our method with a conventional chain-of-thought prompting method (CoT) Wei et al. (2022) and context-aware contrastive decoding method CAD Shi et al. (2023b). We also include the baseline which devises conventional chain-of-thought prompting methods without context (CoT w/o ctx) Wei et al. (2022), to investigate the effect of context in different datasets. Implementation is detailed in the original paper. Notably, compared to the baselines Shi et al. (2023b); Wei et al. (2022), our approach *does not rely on any additional assumptions* and does not require any further processing on the data, to get the performance improvement over these baselines. That is, the inputs of our approach are the same as the baselines Shi et al. (2023b); Wei et al. (2022).

6.6.3 Main Results

Table 6.1 presents evaluation results on the four datasets with three LLM backbone models.

Comparison with Baselines. As we expected, for all LLMs the performance is significantly lower when the context of supporting evidence is absent. Because of the poor performance of the direct prompting method, the context-aware contrastive decoding (CAD) baseline can use its answer distribution as the negative penalty on the positive distribution which is obtained by prompting with both the query and context. However, since the negative answer is not supported by either internal or external knowledge, it can have a more random distribution and limits the effectiveness of contrastive decoding methods. On the other hand, our method DeCoT achieves generally higher improvements on regular CoT compared with CAD by detecting logically incorrect CoTs and penalizing them. Instead of simply contrasting the distributions of positive and negative answers, we use counterfactual context to examine the answer distribution changes, which provides a more fine-grained measurement of the causal effect on LLMs' internal knowledge bias. The consistent performance improvements suggest DeCoT can more accurately detect incorrect CoTs by the measurement, and perform targeted causal intervention.

Logical Reasoning Performance Understanding. We also observe that DeCoT gains relatively better F1 improvements on the SciQ dataset, which reach 18.58%, 46.37% and 14.01% for Flan-T5-xxl, LlaMA-2-7B, and GPT-3.5 respectively. It suggests that accurate logical reasoning paths are more strictly required for scientific questions, and the correctness of the generated CoTs is more crucial. Thus, DeCoT's better performance on the SciQ dataset suggests DeCoT is more effective in debiasing LLMs' logical reasoning ability.

6.6.4 Improving ReAct by DeCoT

Since our main purpose is to find potential spurious correlation and correct reasoning errors with counterfactual debiasing, our approach is compatible with and can improve other model reasoning method variant. To evaluate the generalizability of DeCoT, we apply the proposed

counterfactual reasoning method on ReAct Yao et al. (2023c). Similar to Wei et al. (2022), we use two knowledge-intensive tasks, HotpotQA Yang et al. (2018), a question-answering task and FEVER Thorne et al. (2018), fact verification task. In the evaluation, instead of using context from datasets, ReAct can generate a sequence of reasoning paths with knowledge retrieval queries to find relevant information from external knowledge bases Yao et al. (2023c). From Table 6.2, we observe that by incorporating ReAct generated context as reasoning evidence, DeCoT further improves ReAct's performance by counterfactual reasoning.

Table 6.2: Performance of DeCoT and baselines using retrieved knowledge and reasoning paths from ReAct. Following Yang et al. (2018); Thorne et al. (2018); Yao et al. (2023c), we adopt the EM metric for HotpotQA and the Acc metric for FEVER in the evaluations.

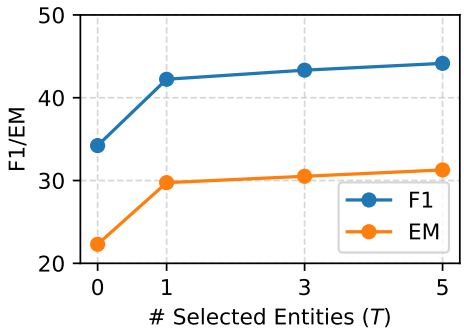
Method	**HotpotQA** EM ↑	**FEVER** Acc ↑
ReAct	26.53	62.63
CoT (w/ ReAct)	20.40	56.12
CAD (w/ ReAct)	27.55	54.02
DeCoT (w/ ReAct)	**36.73**	**66.32**

6.6.5 Impact of the Selected Entities

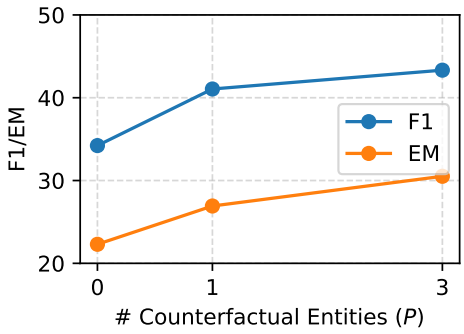
We evaluate the impact of the number of the selected entities on the MuSiQue dataset based on the backbone model GPT-3.5. Since annotations in MuSiQue guarantee the minimal number of chains of thought is 4 hops Ramesh et al. (2023), more factual evidence is required to support the final answer, which makes the impact of the selected entities higher in this case.

To illustrate the trend, we only conduct experiments on a number of the selected entities T with these representative values considering the expensive GPT API costs. In Figure 6.4a, we show the F1 and EM performance w.r.t. the different number of selected entities $T = 0, 1, 3, 5$. We include the result of $T = 0$ which indicates the regular CoT prompting method. With a larger T, it has a higher probability for DeCoT to find more important entities for causal intervention. However, more causal intervention experiments require more counterfactual prompting, which is at the expense of more API calls or inference time. We observe that we can accurately find good

factual entities by selecting the most popular entities.



(a) Impact of T.



(b) Impact of P.

Figure 6.4: Sensitivity studies on the impact of the number of the selected entities and alternative entities. We also include the $T = 0$ and $P = 0$ data points indicating the performance of regular CoT prompting methods. The experiments are conducted on the GPT-3.5 backbone model on the MuSiQue dataset.

6.6.6 Impact of the Alternative Entities

In Figure 6.4b, we also show the performance w.r.t. the different number of alternative entities $P = 0, 1, 3$. For the same reason of computational cost, we only choose a small number of alternative entities. Since the alternative entities are counterfactual, normally the negative causal effect should be similar among the alternative entities. However, practically the alternative entities may further introduce unexpected causal relationships to the context. Thus, by calculating the average causal effect of multiple alternative entities, such unexpected causal effects can be averaged out.

6.6.7 Case Study

We present case study results to understand the impact of context quality on the generated CoTs. In Table 6.3, we show the CoT reasoning paths prompted by DeCoT and the following methods:

- **CoTs w/ Context**: regular CoT prompting with context Wei et al. (2022) (*i.e.,* CoT in Section 6.6.2.).

Table 6.3: Examples of failure CoTs generated from regular CoT prompting, CoT prompting without context, and knowledge post-edited CoT prompting, as well as DeCoT sampled successful CoTs, from four datasets with the GPT3.5 model. In the examples, we highlight factual and logical errors with a red font, while the correct reasoning evidence is in a green font. The edited factuality is also highlighted with a blue font.

Dataset	Question	CoTs w/ Context	CoTs w/o Context	CoT w/ Post-Editing	DeCoT
HotpotQA	The Pluralist school is said to have included what creator of the theory of atomism?	The Pluralist school was a school of pre-Socratic philosophers. The Pluralist school included Leucippus and Democritus. Therefore, the final answer is [Democritus ✗].	One of the key pluralist figures was Greek philosopher Aristotle. Aristotle is known for his theory that nature contains no vacuums. Therefore, the final answer is [Aristotle ✗].	One of the key pluralist figures was Greek philosopher Aristotle. Aristotle is known for his theory of atomism. Therefore, the final answer is [Aristotle ✗].	The Pluralist school is a philosophical movement that emerged in Ancient Greece. It is said to have included Leucippus, the creator of the theory of atomism. Therefore, the final answer is [Leucippus ✓].
MuSiQue	What county contains the city with a radio station that broadcasts to the capital city of the state where the Peace center is located?	The state where the Peace center is located is Greenville, South Carolina. The county containing the city with a radio station broadcasting to Greenville is likely Greenville County. Therefore, the final answer is [Greenville County ✗].	The Peace Center is located in Greenville, South Carolina. The capital city of South Carolina is Columbia. There are two radio stations in Greenville that broadcast to Columbia - WJMZ-FM and WMYI. WMYI, which is located in Greenville, broadcasts to the capital city of Columbia. Therefore, the final answer is [Greenville County ✗].	The Peace Center is located in Greenville, South Carolina. The capital city of South Carolina is Columbia. There are two radio stations in Greenville that broadcast to Columbia - WJMZ-FM and WMYI. WMYI is located in Hendersonville, North Carolina. Therefore, the final answer is [Hendersonville ✗].	The Peace Center is located in Greenville, South Carolina. The radio station WWNQ is licensed to Forest Acres and broadcasts to Columbia, which is the capital city of South Carolina. Forest Acres is a city in Richland County, South Carolina. Therefore, the final answer is [Richland County ✓].
SciQ	By exploding what the space probes get separated from their launchers?	The context is discussing how space probes use exploding bolts to separate from their launchers. The question is asking what the space probes are separated from during this process. Therefore, the final answer is [their launchers ✗].	When a space probe is launched, it is attached to a rocket or launcher. Once the rocket has reached a certain point in space, it explodes, releasing the space probe. Therefore, the final answer is [the rocket or launcher ✗].	When a space probe is launched, it is attached to a rocket or launcher. ~~Once the rocket has reached a certain point in space, it explodes, releasing the space probe.~~ Therefore, the final answer is [the explosion or detonation ✗].	The context states that space probes may be separated from their launchers by exploding bolts. Exploding bolts are bolts that explode or break apart in order to separate two objects. Therefore, the final answer is [bolts ✓].
WikiHop	What's the genre of Terra Brasilis?	"Genre" refers to a category of artistic work. Terra Brasilis is a famous album by Brazilian musician Milton Nascimento. Therefore, the final answer is [Brazilian music ✗].	"Terra Brasilis" is a song by Brazilian composer Caetano Veloso. The song is from his 1978 album "Muito (Dentro Da Estrela Azulada)". Therefore, the final answer is [Brazilian ✗].	"Terra Brasilis" is an album by Brazilian composer Antônio Carlos Jobim. ~~The song is from his 1978 album "Muito (Dentro Da Estrela Azulada)".~~ Therefore, the final answer is [Brazilian ✗].	The album Terra Brasilis was released in 1980 by Brazilian composer Antônio Carlos Jobim. Jobim is known for his contributions to bossa nova music. Therefore, the final answer is [bossa nova ✓].

- **CoTs w/o Context**: regular CoT prompting with only task queries (*i.e.,* CoT w/o ctx in Section 6.6.2.).
- **CoT w/ Post-Editing**: CoT with knowledge post-edit Zhao et al. (2023b), when factual contradictions exist in the context.

For all the methods, the CoTs are prompted from the backbone GPT-3.5 model.

Case Study of CoTs with Post-Editing. We observe that CoTs generated without context are likely to contain incorrect knowledge, which can further mislead the reasoning paths. For example, the directly generated CoTs of the question in the WikiHop dataset say "Terra Brasilis is a song by Caetano Veloso", which is factually incorrect (highlighted in a red font). Due to this incorrect assumption made from LLMs' hallucination, the following reasoning paths are misled to talk about irrelevant information (*e.g.*, "Caetano Veloso's album"), and thus the answer is wrong even with factual edit (highlighted in a blue font).

Case Study of CoTs with Context. Compared to CoTs generated without context, we observe that the CoTs prompted with context can be factually more faithful. However, the logical reasoning of these CoTs can still be wrong. For example, the CoTs generated with the context in the HotpotQA dataset correctly locate "Leucippus and Democritus" as the two "Pluralist school" members (highlighted in a red font). However, instead of answering with the one who created the school, the LLM mistakenly chooses the wrong answer "Democritus". We highlight the correct reasoning paths in green font to show that the key point in answering this question is by identifying the "creator".

6.7 Chapter Summary

In this paper, we formally examine the LLM's internal knowledge bias and identify it as the confounder by a structural causal model (SCM). We revisit the irrelevant contexts issue and further discover the spurious correlation between the LLM and task queries, which can further impair the LLM's CoT reasoning abilities. Then, we propose DeCoT, a debiasing chain-of-thought prompting method in knowledge-intensive tasks, which alleviates the spurious correlation and enables

the LLM to find more accurate and logically sound responses. Following the previous evaluation setting Welbl et al. (2017); Trivedi et al. (2022b); Shi et al. (2023b); Wei et al. (2022), and using the same inputs as the baselines, extensive experimental results and case studies validate the effectiveness of our method DeCoT.

This chapter, in full, is a reprint of the material as it appears in "DeCoT: Debiasing Chain-of-Thought for Knowledge-Intensive Tasks in Large Language Models via Causal Intervention" by Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan A. Rossi, Sungchul Kim, Anup Rao, and Julian McAuley, published in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. The dissertation author was the primary investigator and author of this paper.

Chapter 7

Chain-of-Thought Reasoning via Latent State Transition (CTRLS)

7.1 Introduction

Chain-of-thought (CoT) reasoning has emerged as an effective paradigm for enabling large language models (LLMs) to tackle complex tasks by decomposing them into structured, interpretable intermediate reasoning steps (Wei et al., 2022; Kojima et al., 2022; Wu et al., 2024d). However, conventional CoT prompting lacks transition-aware modelling, such that each step is generated autoregressively without capturing the underlying dynamics of reasoning transitions, limiting explainable exploration and diversity Yu et al. (2025); Zhang et al. (2025b); Gan et al. (2025); Hou et al. (2023), in reasoning trajectories (Wei et al., 2022; Kveton et al., 2025; Xia et al., 2025a). On the other end, structured reasoning frameworks (e.g., Toolchain* (Zhuang et al.), program synthesis (Zhang et al., 2023c), knowledge graph (Wu et al., 2025b)) enforce rigid intermediate steps via API calls or logic traces, offering structural control but sacrificing flexibility and semantic adaptability (Kojima et al., 2022; Wu et al., 2025b).

To illustrate the limitations of conventional CoT prompting, Figure 7.1 contrasts standard autoregressive reasoning with our transition-aware framework. On the left, traditional CoT unfolds step-by-step without modelling transitions between reasoning steps, often leading to prema-

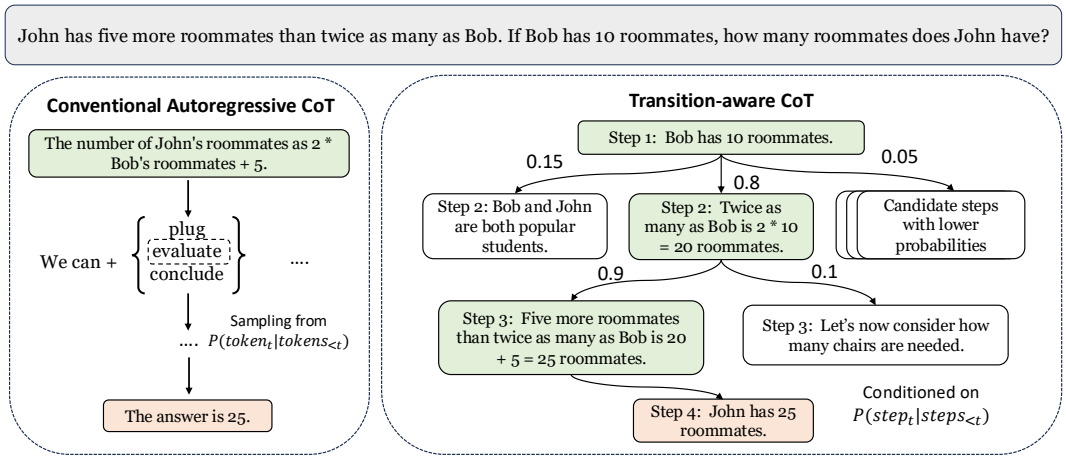


Figure 7.1: Illustration of the difference between conventional CoT prompting and CTRLS.

ture commitment and limited trajectory diversity. In contrast, our method models reasoning as a stochastic trajectory over latent states, enabling explicit transition modelling and exploration of alternative reasoning paths. This comparison highlights the potential of transition-aware CoT to uncover more effective and diverse reasoning strategies.

To bridge the extremes, we propose transition-aware CoT reasoning `CTRLS`, a novel perspective that frames reasoning as structured trajectories within a latent state space. Each reasoning step corresponds to a continuous latent semantic state, with transitions dynamically learned via a latent-state MDP formulation. This modelling explicitly captures reasoning regularities and semantic relationships between intermediate steps, supporting explainable exploration.

Modelling transition-aware CoT reasoning presents fundamental challenges: (i) the need to infer latent semantic states that capture the progression of reasoning steps, despite the lack of explicit supervision; (ii) learning stable and generalizable transition dynamics across these latent states; and (iii) adapting LLMs to generate coherent reasoning steps conditioned on these latent abstractions. To address these challenges holistically, we propose a unified variational model that jointly learns a latent encoder, a transition policy, and a reasoning adapter for the LLM, optimized under a single **evidence lower bound (ELBO)** objective. This structured design enables semantic grounding and modular reasoning, and naturally supports distributional reinforcement learning (Bellemare et al., 2017; Dabney et al., 2018; Wu et al., 2022b) by treating reasoning actions as

stochastic policies over latent states. To ensure robust exploration and avoid degenerate trajectories, CTRLS incorporates on-policy optimization with entropy regularization and epsilon-greedy sampling. Overall, our framework provides a principled and tractable foundation for controllable, structured, and adaptive CoT reasoning. We summarize our contributions as follows:

- We introduce a latent-space MDP formulation for explicit modelling of chain-of-thought transitions.
- We propose a distributional reinforcement learning method tailored for robust and explainable exploration in CoT generation.
- We theoretically derive finite-sample evidence lower bounds (ELBO) as the learning objective of CTRLS pre-training.
- We demonstrate empirical improvements in reasoning accuracy, diversity, and exploration efficiency, and showcase enhanced explainability on standard benchmarks.

7.2 Related Work

Chain-of-Thought Reasoning. Chain-of-Thought (CoT) prompting improves large language models (LLMs) by encouraging intermediate reasoning steps before producing final answers (Wei et al., 2022; Chu et al., 2023b; Wu et al., 2024d; Wang et al., 2026a). To enhance reasoning quality, prior works introduce self-evaluation (Ling et al., 2023; Shinn et al., 2023) or integrate external knowledge (Zhao et al., 2023b). Coconut (Hao et al., 2024) explores reasoning in continuous latent space to bypass token-level constraints. In contrast, we focus on controllability and interpretability via structured latent transitions in language space. Beyond linear CoT, recent work explores structured reasoning, including Tree-of-Thought (ToT) (Yao et al., 2023b) and Chain-of-Preference Optimization (CPO) (Zhang et al., 2024c), which guide CoT using preferred ToT trajectories. Building on these insights, we model reasoning as transitions over a latent state space, enabling structured stepwise guidance without explicit search.

Reinforcement Learning for LLM Reasoning. Reinforcement learning (RL) has been widely

used to enhance the reasoning abilities of large language models (LLMs), particularly via Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a) and Direct Preference Optimization (DPO) (Rafailov et al., 2023; Li et al., 2025c; Wu et al., 2025c; Huang et al., 2025e, 2026a), which learns reward models from human preferences and optimizes LLMs using policy gradient methods such as PPO. While effective for preference alignment, RL-based reasoning faces challenges from sparse and delayed rewards. To address this, recent works introduce outcome-based rewards (Cobbe et al., 2021; Uesato et al., 2022) or process-based feedback on intermediate steps (Lightman et al., 2023; Choudhury, 2025), sometimes leveraging verifiers (Su et al., 2025; Mroueh, 2025; Yue et al., 2025; Mundada et al., 2026) (RLVR) distilled from GPT-4 (Zhang et al., 2024c) and off-policy evaluation (Huang et al., 2025c). In contrast, we directly model reasoning dynamics in a latent state space and fine-tune transition behaviours using policy gradients, without relying on external reward models. This enables scalable and interpretable optimization of reasoning trajectories in a self-contained framework.

7.3 Preliminaries

7.3.1 Chain-of-thought Reasoning

Chain-of-thought (CoT) reasoning sequentially generates step-by-step intermediate reasoning toward solving complex tasks by large language models (LLMs) (Wei et al., 2022; Kojima et al., 2022). Given an initial prompt or query x_0, the LLM policy μ , iteratively generates reasoning steps $x = (x_1, x_2, \ldots, x_T)$, culminating in a final answer prediction y. Each reasoning step x_t consists of sequentially sampled tokens conditioned on all previously generated tokens:

$$x_t \sim \mu(\cdot|x_0, x_{<t}), \quad y \sim \mu(\cdot|x_0, x), \tag{7.1}$$

where x_0 is the input query or prompt (Wei et al., 2022; Wu et al., 2025b; Li et al., 2025d). This autoregressive nature of CoT reasoning presents inherent challenges for controllable gener-

ation, as decisions at each step significantly impact subsequent reasoning trajectories. Instead of explicitly modelling this sequential token generation process, we consider reasoning via latent state-transition, where we assume a latent state encoded by an *thought abstraction* model $S_t = \rho_\phi(x_{<t})$ and a state transition process $p_\theta(S_{t+1}|S_t)$.

Latent spaces offer a path to more *explainable* decision making. We take a task-driven view in which intermediate steps are semantically meaningful and form transparent trajectories. Following Yu et al. (2025); Zhang et al. (2025b); Gan et al. (2025); Hou et al. (2023) on steps to human rationale alignment, we further focus on *transition dynamics* among latent reasoning states. Explicitly modeling these transitions makes the process explainable and aligns with our formulation of state-based reasoning and controllable exploration (detailed in Section 7.3.3).

7.3.2 Distributional Reinforcement Learning

Distributional Reinforcement Learning (DRL) explicitly models uncertainty by representing actions as probability distributions rather than deterministic or discrete selections (Bellemare et al., 2017; Dabney et al., 2018). Specifically, we parametrise the policy π_θ to output Dirichlet distributions over potential next reasoning step in CoTs, which naturally express uncertainty and enable richer exploratory behaviour (Chou et al., 2017). Given state s_t, action distributions a_t are drawn as:

$$a_t \sim \pi_\theta(\cdot|s_t), \quad a_t \in \Delta(\mathcal{A}). \tag{7.2}$$

This formulation inherently captures the uncertainty of stochastic actions over the transition probability distribution of a chain of thoughts. The distributional Bellman operator explicitly incorporates uncertainty into state transitions:

$$\mathcal{T}Z(s_t, a_t) \overset{D}{=} R(s_t, a_t) + \gamma Z(s_{t+1}, a_{t+1}), \qquad a_{t+1} \sim \pi_\theta(\cdot|s_{t+1}) \tag{7.3}$$

where $\overset{D}{=}$ denotes distributional equality and $Z(s, a)$ is the return distribution from state-action pairs. Thus, we could enable efficient exploration over CoT transition distribution, essential for improved

controllability and explainability in downstream tasks (Haarnoja et al., 2018; Lillicrap et al., 2015).

7.3.3 Formulation: Transition-aware Chain-of-Thought as an MDP

We cast the chain-of-thought (CoT) decoding process into the standard Markov Decision Process (MDP) framework (Bai et al., 2025; Zhang et al., 2025a; Wu et al., 2025b), defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$. This allows us to leverage reinforcement learning algorithms to optimise reasoning trajectories in LLMs.

State and Transition Dynamics At each time step t, the agent observes a latent reasoning state $s_t \in \mathcal{S}$ that captures the semantic context of the prompt and all previously generated reasoning steps. Formally, we obtain $s_t \sim \rho_\phi(\cdot \mid x_0, x_{<t})$ via a stochastic abstraction model $\rho_\phi : \mathcal{X}_{<t} \to \Delta(\mathcal{S})$, where $\mathcal{X}$ is the space of initial prompts and the sequence of reasoning prefixes (detailed in Section 7.4.2). Here, x_0 is the input query or prompt and $x_{<t} = (x_1, \ldots, x_{t-1})$ are the previously generated reasoning segments. We model latent state evolution with a learned transition matrix $P(s_{t+1} \mid s_t, a_t)$. This kernel encapsulates how one reasoning segment influences subsequent latent abstractions and can be trained jointly with the policy.

Distributional Action Sampling Our formulation distinctly leverages a distributional perspective for action sampling within the action space $\mathcal{A}$, which encompasses distributions over admissible reasoning segments rather than discrete or deterministic selections. At each latent reasoning state s_t, the policy π_θ explicitly outputs a probability distribution, capturing the inherent epistemic uncertainty in selecting reasoning segments. This distributional approach facilitates richer exploratory behaviour by enabling the policy to represent a spectrum of plausible reasoning steps, each sampled according to a learned Dirichlet distribution:

$$a_t \sim \pi_\theta(\cdot \mid s_t), \quad a_t \in \Delta(\mathcal{A}), \quad x_t \sim P_\omega(\cdot \mid a_t, x_{<t}),$$

where $\Delta(\mathcal{A})$ denotes the simplex representing the space of action distributions. Here, θ parametrises a policy network designed to output Dirichlet parameters, thus explicitly modelling uncertainty and

providing principled exploration strategies in the CoT action space. P_ω denotes the adapted LLM with adapted parameters introduced in the backbone LLM μ for action-conditioned generation, which is explained in detail in Section 7.4.3.

Trajectory-level Reward Aligning with the formulation of LLMs, the quality of the CoT $\tau = (x_1, \ldots, x_T)$ is evaluated by the final answer y. Let $y^\star$ denote the ground-truth answer associated with the query x_0. The episodic reward is therefore binary $R(\tau, x_0) = \mathbf{1}\{y = y^*\}$. Our learning objective is to maximise the expected terminal accuracy under the controllable policy

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot|x_0)}\big[R(\tau, x_0)\big], \tag{7.4}$$

where $R(\tau)$ is sparse and unbiased, corresponding to the accuracy reported in the evaluation.

7.4 CTRLS: Chain-of-Thought Reasoning via Latent State Transition

Modeling transition-aware CoT reasoning presents unique challenges: it requires learning latent state representations to capture the semantic progression of reasoning steps, modeling transitions that reflect meaningful and generalizable reasoning dynamics, and conditioning the backbone LLM to generate coherent next steps guided by these latent states. To address these, CTRLS adopts a unified variational framework and an MDP perspective, implementing three core components: (i) a stochastic encoder that abstracts reasoning into latent states via an inference model (Section 7.4.2), (ii) a state-conditioned UNet that injects latent guidance into token representations and models transitions through a policy network (Section 7.4.3), and (iii) a two-phase alignment and fine-tuning scheme that combines online pre-training with on-policy reinforcement learning (Section 7.4.4) (illustrated in Figure 7.2). This structure supports transition-aware reasoning by optimizing a unified ELBO objective, explicitly modeling uncertainty for principled exploration and controllable, structured generation.

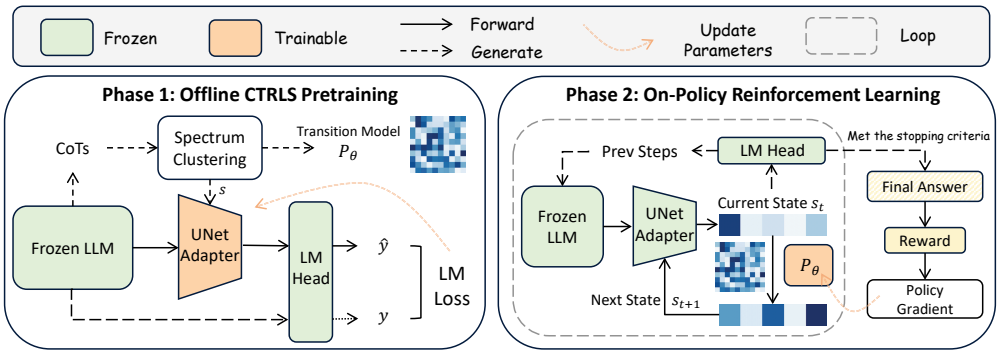


Figure 7.2: An overview of the proposed two-phase alignment and fine-tuning scheme.

7.4.1 Mathematical Assumptions

Assumption 5.1 First-order Markov Assumption: The latent reasoning states follow a first-order Markov process. That is, the transition probability depends only on the previous latent state:

$$P_\theta(z_t \mid x_{<t}, z_{<t}) = P_\theta(s_t \mid s_{t-1})$$

This assumption simplifies the transition dynamics and is justified by the fact that each latent state encodes the full reasoning prefix.

Assumption 5.2 Autoregressive Factorization of the Variational Posterior: The variational posterior over latent states is factorized autoregressively to align with the left-to-right generation of LLMs:

$$Q_\phi(z_{1:T} \mid x_{1:T}) = \prod_{t=1}^{T} Q_\phi(z_t \mid x_{\leq t})$$

This ensures consistency between the inference model and the sequential nature of language generation.

7.4.2 Latent State Encoding

Based on our MDP formulation in Section 7.3.3, we introduce a variational approximation $Q_\phi(z_{1:T}|x_{1:T})$, parametrised by ϕ, encoding the prompt x_0 and reasoning steps $x_{<t}$ into latent state representations.

Definition 7.1. Inference Model via Variational Posterior We introduce a variational approximation $Q_\phi(z_{1:T} \mid x_{1:T})$. Consistent with the sequential nature of the data, it factorises autoregressively:

$$Q_\phi(z_{1:T} \mid x_{1:T}) = \prod_{t=1}^{T} Q_\phi\big(z_t \mid x_{\leq t}\big).$$

For each time step t, the conditional model $Q_\phi\big(z_t \mid x_{\leq t}\big)$ outputs a probability distribution over the N possible values of z_t. It can be instantiated by a Gaussian-mixture posterior, a linear classifier, or a multilayer perceptron (MLP) applied to the hidden representation produced by an encoder.

The autoregressive factorization mirrors the temporal ordering of the observations, enabling the posterior at step t to depend only on the current and past inputs $x_{\leq t}$. By sampling z_t rather than deterministically encoding it, the model captures both semantic content and epistemic uncertainty, reflecting variability in how the system "thinks" before emitting each step.

To instantiate this, we follow (Kveton et al., 2025) to extract token representation $E_t \in \mathbb{R}^{n_t \times d}$ and compute the Gram matrix $G_t = E_t^\top E_t$. Then, to capture the reasoning semantics, we perform spectrum decomposition and take the flattened most informative top k eigenspaces as the representation,

$$e_t := \big[\sqrt{\lambda_1} \cdot \mathbf{q}_1; \ldots; \sqrt{\lambda_k} \cdot \mathbf{q}_k\big] \in \mathbb{R}^{kd}, \tag{7.5}$$

where $\mathbf{q}_k$ are the corresponding eigenvectors. We apply k-means clustering to all such e_t and each reasoning state $s_{t-1} \sim \rho_\phi(\cdot \mid x_{<t})$ is then assigned as the probability distribution corresponding to clustering centroids $\{\gamma_j\}_{j=1}^K$, such that $z_t = \sum_{j=1}^K s_{t,j} \cdot \gamma_j$. This continuous relaxation Q_ϕ enables structured reasoning over the latent state space and supports subsequent transition modelling.

7.4.3 State-aware Chain-of-thought Alignment

Definition 7.2. State-aware Generative Model Let $\{(x_t, z_t)\}_{t=1}^T$ denote a sequence of observed variables $x_{1:T} \in \mathcal{X}^T$ and latent states $z_{1:T} \in \mathcal{Z}^T$. For parameters ω (emission) and θ (transition),

the joint distribution factorises autoregressively as

$$P_{\omega,\theta}(x_{1:T}, z_{1:T}) = \prod_{t=1}^{T} P_{\omega,\theta}\big(x_t, z_t \mid x_{<t}, z_{<t}\big),$$

where each factor decomposes into a transition term and an emission term

$$P_{\omega,\theta}\big(x_t, z_t \mid x_{<t}, z_{<t}\big) = P_\theta\big(z_t \mid x_{<t}, z_{<t}\big) P_\omega\big(x_t \mid x_{<t}, z_{\leq t}\big). \tag{7.6}$$

The transition model P_θ places a prior on the next latent state; it may depend on past observations but is often simplified by the first-order Markov assumption $P_\theta(z_t \mid z_{t-1})$. The emission model P_ω generates the current observation conditioned on the complete latent trajectory up to time t. Together, P_θ and P_ω fully specify the stochastic process underlying the data.

Transition Model P_θ To model the conditional latent transition, we first encode the previous reasoning steps into a latent state distribution $s_{t-1} \sim \rho_\phi(\cdot \mid x_{<t})$, as described in Section 7.4.2. Given that the latent representation z_t is deterministically constructed via $z_t = \sum_{j=1}^{K} s_{t,j}\gamma_j$ (see Section 7.4.2), the transition probability naturally reduces from transitions between latent states z_t to transitions between their underlying distributions s_t. Specifically, applying the Markov property, we have

$$P_\theta(z_t \mid x_{<t}, z_{<t}) = P_\theta(s_t \mid s_{<t}) = P_\theta(s_t \mid s_{t-1}), \tag{7.7}$$

where the first equality leverages the deterministic relationship between z_t and s_t, and the second equality explicitly invokes the Markov assumption (Wu et al., 2022b; Eysenbach et al., 2020).

Generation Modelling P_ω To enable state-aware LLM generation, we design a transformation module that reshapes each token's last hidden state representation $E_{t,i} \in \mathbb{R}^d$ conditioned on a step-wise latent reasoning state $s_t \in \mathbb{R}^{d'}$ based on MDP transition. Specifically, according to the Markov

property

$$
\begin{aligned}
&E'_{t,i} = \mathcal{F}_{\omega_u}\left([\mathcal{F}_{\omega_d}(E_{t,i}); z_t]\right), \quad i = 1, 2, \cdots, n_t \\
&P_\omega(x_t \mid z_{<t}, x_{<t}) = P_\omega(x_t \mid z_{t-1}, x_{<t}) = \mu(x_t \mid [E'_{1,i<n_1}; \cdots; E'_{t-1,i<n_{t-1}}], \omega),
\end{aligned} \tag{7.8}
$$

where $\omega = [\omega_d; \omega_u]$ is a U-Net module that encodes token representation $E_{t,i}$ into a low-rank latent, projects and fuses z_t via a bottleneck interaction, and decodes the result back to dimension d. This conditional encoder allows the model to dynamically adjust token representations in sync with the evolving state, with only $O(r(d + d'))$ additional parameters.

Thus, the one-step inference of state-conditional generation in (7.6) is realised as

$$
\begin{aligned}
P_{\omega,\theta}\left(x_t, z_t \mid x_{<t}, z_{<t}\right) &= P_\omega(x_t \mid z_{<t}, x_{<t})\ P_\theta(z_t \mid x_{<t}, z_{<t}) \\
&= \mu(x_t \mid H_{t-1}, \omega)\ P_\theta(s_t \mid s_{t-1}), \qquad (7.9) \\
H_{t-1} &\coloneqq \left[E'_{1,\,i<n_1};\ \cdots;\ E'_{t-1,\,i<n_{t-1}}\right]. \qquad (7.10)
\end{aligned}
$$

where such factorization makes clear how latent dynamics and token sampling interlock. Then we formally introduce the evidence lower bound (ELBO) of the proposed variational model as the objective for model pre-training.

Theorem 7.3 (Evidence Lower-Bound)**.** *Consider a latent-state generative model with joint density* $P_{\omega,\theta}(x_{1:T}, z_{1:T}) = \prod_{t=1}^{T} P_\omega\big(x_t \mid x_{<t}, z_{\le t}\big) P_\theta\big(z_t \mid x_{<t}, z_{<t}\big)$*, and let* $Q_\phi(z_{1:T} \mid x_{1:T}) = \prod_{t=1}^{T} Q_\phi\big(z_t \mid x_{\le t}\big)$ *be any variational distribution. Then, for the learnable parameters* ω, θ, ϕ*, the marginal log-likelihood of the observations admits the lower bound*

$$
\mathcal{L}(\omega, \theta, \phi) = \sum_{t=1}^{T} \mathbb{E}_{Q_\phi(z_{\le t} \mid x_{\le t})}\left[\log P_\omega(x_t \mid x_{<t}, z_{\le t})\right] \tag{7.11}
$$

$$
- \sum_{t=1}^{T} \mathbb{E}_{Q_\phi(z_{<t} \mid x_{\le t-1})}\Big[\mathrm{KL}\big(Q_\phi(z_t \mid x_{\le t}) \,\|\, P_\theta(z_t \mid x_{<t}, z_{<t})\big)\Big] \tag{7.12}
$$

Equality holds if and only if $Q_\phi(z_{1:T} \mid x_{1:T}) = P_\theta(z_{1:T} \mid x_{1:T})$*, i.e. when the variational posterior*

matches the true posterior. Maximising $\mathcal{L}$ therefore constitutes a tractable surrogate objective whose optimisation w.r.t ω, θ, ϕ simultaneously (i) maximises the data likelihood and (ii) minimises the posterior gap.

Theorem 5.3 shows that the ELBO provides a tractable way to align latent transitions with reasoning steps. In practice, this means our pretraining objective is not only theoretically sound but also effective: as seen on GSM8K and MATH, optimizing under this objective leads to higher exploration accuracy and fewer spurious steps (detailed later in Section 7.5.2).

7.4.4 On-Policy Chain-of-thought Reinforcement Learning

After offline pre-training, we further enable on-policy reinforcement learning that optimises only the state-transition model P_θ through trajectories generated by the current policy. At each decision step t, we sample an action distribution $a_t \sim \pi_\theta(\cdot \mid s_t)$ conditioned on the current latent state s_t, and subsequently generate the reasoning segment x_t through the state-conditioned LLM generation P_ω. Iteratively repeating this process yields complete trajectories $\tau = (s_t, a_t, s_{t+1})_{t=1}^{T}$ ending with a final predicted answer y.

Trajectory-level Reward and Bellman Function We evaluate each trajectory τ based on the correctness of its final answer y against the ground truth y^*, providing a binary episodic reward:

$$R(\tau, x_0) = \mathbf{1}\{y = y^*\}. \tag{7.13}$$

We adopt the distributional Bellman operator (7.3) to model uncertainty explicitly in state transitions and returns.

Exploration via Epsilon-Greedy To effectively balance exploration and exploitation, we implement epsilon-greedy exploration. Specifically, the exploration-enabled action distribution is obtained by mixing the learned Dirichlet distribution with a uniform distribution over actions:

$$\tilde{\pi}_\theta(a|s_t) = (1-\epsilon)\pi_\theta(a|s_t) + \epsilon \text{Uniform}(\mathcal{A}), \tag{7.14}$$

where $\epsilon \in [0, 1]$ controls the exploration-exploitation trade-off. With probability ϵ, actions are thus sampled uniformly from the action simplex, promoting exploration of diverse reasoning segments, while with probability $1 - \epsilon$, actions follow the learned Dirichlet distribution, ensuring exploitation of promising behaviours.

Entropy Regularization To further encourage robust exploration and prevent premature convergence to suboptimal solutions, we employ entropy regularization. By adding an entropy-based penalty to the learning objective, the policy maintains diversity in the action distributions, effectively exploring the full action space and discovering potentially more rewarding trajectories. The entropy term is defined as:

$$\mathcal{H}(\pi_\theta(s_t)) = -\sum_{a \in \mathcal{A}} \pi_\theta(a|s_t) \log \pi_\theta(a|s_t). \tag{7.15}$$

Overall Learning Objective and Policy Gradient Integrating trajectory rewards and exploration strategies, our reinforcement learning objective maximises expected trajectory rewards augmented with entropy-based exploration:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\cdot|x_0)} \Big[R(\tau, x_0) + \alpha \sum_{t=1}^{T} \mathcal{H}(\pi_\theta(s_t)) \Big], \tag{7.16}$$

where α controls the strength of entropy regularization. Using the REINFORCE estimator, the gradient of this objective with respect to policy parameters θ is computed as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \Big[\big(R(\tau) + \alpha \sum_{t=1}^{T} \mathcal{H}(\pi_\theta(s_t)) \big) \times \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \Big]. \tag{7.17}$$

This gradient formulation aligns with accuracy-based evaluation metrics in CoT benchmarks, guiding the optimization of the distributional state-transition model. The on-policy reinforcement learning procedure is detailed in the original paper.

7.5 Experiments

7.5.1 Experimental Setup

We conduct experiments on two instruction-tuned language models, LLaMA-3.2-3B-Instruct (Grattafiori et al., 2024) and Qwen2.5-3B-Instruct (Team, 2024b). Both models serve as the backbone for integrating our framework without modifying any pretrained weights. We evaluate on two math reasoning benchmarks, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021), which might require step-by-step CoT generation to solve arithmetic and competition-level mathematics problems, respectively. Implementation details are in the original paper.

Table 7.1: Comparison of exploration accuracy and success rate of CTRLS and the base models.

Generation Temperature	Exploration	LlaMA3.2				Qwen2.5			
		GSM8K		MATH		GSM8K		MATH	
		pass@20	Succ.	pass@20	Succ.	pass@20	Succ.	pass@20	Succ.
$\eta = 0.5$	Base	72.5	95.0	50.0	90.0	87.5	**100.0**	**65.0**	65.0
	$\epsilon = 0.1$	77.5	**100.0**	47.5	**95.0**	**90.0**	**100.0**	57.5	60.0
	$\epsilon = 0.3$	75.0	**100.0**	**60.0**	**95.0**	87.5	**100.0**	62.5	52.5
	$\epsilon = 0.5$	**82.5**	**100.0**	52.5	90.0	87.5	**100.0**	**65.0**	**75.0**
$\eta = 0.7$	Base	77.5	**100.0**	47.5	**97.5**	80.0	**100.0**	72.5	**67.5**
	$\epsilon = 0.1$	77.5	**100.0**	50.0	**97.5**	**85.0**	97.5	62.5	65.0
	$\epsilon = 0.3$	70.0	**100.0**	**60.0**	**97.5**	82.5	97.5	62.5	**67.5**
	$\epsilon = 0.5$	**85.0**	**100.0**	47.5	90.0	80.0	97.5	**77.5**	**67.5**

7.5.2 State Transition-aware Exploration

To assess the controllability of our pre-trained, transition-aware chain-of-thought generator, we compare CTRLS against the corresponding base models (LlaMA3.2 and Qwen2.5). In Table 7.1, we report results for two generation temperatures, $\eta \in \{0.5, 0.7\}$, and an ϵ-greedy exploration strategy with $\epsilon \in \{0.1, 0.3, 0.5\}$. For each test question and each (η, ϵ) configuration, we sample 20 reasoning trajectories from both the base model and CTRLS. We measure **exploration accuracy (pass@20)** as the fraction of questions for which the correct answer appears in at least one of the 20 samples, and **success rate (Succ.)** as the proportion of samples that yield a valid final answer after chain-of-thought generation.

Based on the observations in Table 7.1, CTRLS consistently outperforms its backbone counterparts in both exploration accuracy and success rate across all datasets. Although performance varies with the choice of ϵ, CTRLS surpasses purely temperature-based sampling in every setting, confirming that explicit state-transition guidance yields more effective exploratory behaviour. Such effective control over latent space exploration is consistent with the theoretical guarantees established in Theorem 7.3. The more informative trajectories produced by CTRLS provide a stronger learning signal for subsequent reinforcement learning, accelerating policy improvement in later training stages.

7.5.3 Impact of Exploration Configurations

In Table 7.2, we experiment on different exploration configurations and study the effects for on-policy reinforcement learning. **Entropy regularisation.** We observe that raising the entropy weight from $H{=}0$ to $H{=}0.01$ consistently broadens the search space. For **LLaMA-3.2**, this translates into a modest but steady gain on both datasets, while **Qwen2.5** shows an even clearer trend, where $H{=}0.01$ delivers the best accuracy on both datasets, with a minor loss in success rate. These results confirm that a small amount of entropy pressure prevents premature convergence to high-probability but sub-optimal transitions.

Table 7.2: RL performance under entropy regularization (left block) and ϵ-greedy exploration (right block).

Model	Entropy	GSM8K		MATH		ϵ-Greedy	GSM8K		MATH	
		pass@20	Succ.	pass@20	Succ.		pass@20	Succ.	pass@20	Succ.
LlaMA3.2	Base	55.6	80.0	41.0	71.8	Base	55.6	80.0	41.0	71.8
	$H = 0.0$	57.6	80.8	42.8	72.2	$\epsilon = 0.1$	57.4	80.4	43.0	73.4
	$H = 0.01$	56.8	79.6	42.6	74.0	$\epsilon = 0.3$	55.8	79.4	42.0	71.2
Qwen2.5	Base	64.2	91.8	22.6	33.8	Base	64.2	91.8	22.6	33.8
	$H = 0.0$	65.8	89.6	22.8	33.4	$\epsilon = 0.1$	66.2	89.8	22.2	32.2
	$H = 0.01$	66.4	90.2	23.6	34.6	$\epsilon = 0.3$	66.0	90.2	23.4	34.2

ϵ-greedy sampling. Injecting stochastic jumps at every step also fosters diversity, but the magnitude of ϵ matters. For **LLaMA-3.2**, a mild setting ($\epsilon{=}0.1$) yields the highest overall gains, while a larger perturbation ($\epsilon{=}0.3$) instead hurts both accuracy and solution validity, echoing the classic

explore–exploit trade-off. **Qwen2.5** is more robust for both ϵ values, yet excessive randomisation still reduces the success rate on the harder MATH dataset.

Question: For how many positive integers n is $n^2 - 3n + 2$ a prime number?

Base (Qwen2.5)

Final: $\boxed{2}$

Steps:

1. First, we factorize $n^2 - 3n + 2$. This can be written as $(n-1)(n-2)$.
2. For $(n-1)(n-2)$ to be a prime number, one of the factors must be 1 and the other must be the prime number itself.
3. Therefore, for $(n-1)(n-2)$ to be prime, either $n-1=1$ or $n-2=1$ must hold true.
4. (Wrong conclusion) Thus, there are two values of n that satisfy the condition: $n=2$ and $n=3$.

CTRLS (Qwen2.5)

Final: $\boxed{1}$

Steps:

1. (Same factorization) First, we factorize $n^2 - 3n + 2$. This can be written as $(n-1)(n-2)$.
2. (Same condition) For $(n-1)(n-2)$ to be a prime number, one of the factors must be equal to 1 because a prime number has exactly two distinct positive divisors: 1 and itself.
3. (Same derivation) Therefore, either $n-1=1$ or $n-2=1$..
4. (Self-reflection) However, we need to check if these values satisfy the condition that $(n-1)(n-2)$ is indeed a prime number.
5. (reduce wrong answer) When $n=2$, $(n-1)(n-2)=(2-1)(2-2)=1\cdot 0=0$, which is not a prime number.
6. (Correct conclusion) Thus, there is only one value of n that satisfies the condition, namely $n=3$.

Figure 7.3: **Qualitative comparison.** CTRLS correctly verifies candidate solutions and filters out invalid cases, while the baseline fails to check whether the resulting value is truly prime.

Table 7.3: GPT-4 rubric scores (0–10) on GSM8K with Qwen. Averaged over 100 samples per setting. We evaluate on four different aspects: (**S**)elf-reflection, (**A**)lgebra, (**C**)oherence, (**H**)allucination.

Setting	S	A	C	H↓	Overall
Base	7.62	8.44	8.23	8.51	8.05
$\epsilon = 0.0, H = 0.0$	7.82	8.83	8.43	8.66	8.34
$\epsilon = 0.1, H = 0.0$	7.74	8.65	8.33	8.51	8.17
$\epsilon = 0.1, H = 0.01$	7.73	8.54	8.35	8.55	8.16
$\epsilon = 0.3, H = 0.0$	7.80	8.77	8.45	8.68	8.30
$\epsilon = 0.3, H = 0.01$	7.72	8.73	8.37	8.60	8.20

7.5.4 Case Study

Self-reflection One notable advantage of our transition-based reasoning is its ability to perform self-reflection. As illustrated in Figure 7.3, both the baseline and CTRLS derive the correct candidate values for n such that $n^2 - 3n + 2$ is prime. However, the baseline prematurely outputs both

values as correct without verifying whether the resulting expression is actually a prime number. In contrast, CTRLS continues reasoning by explicitly validating the primality condition for each candidate. It correctly identifies $n = 2$ as invalid, as $(2-1)(2-2) = 1 \cdot 0 = 0$, which is not a prime. This self-reflective validation step leads to the correct final answer. Such state-aware step transitions, which is semantically grounded on explainable states, help avoid early commitment to incorrect conclusions.

Corrected algebra errors and reduced hallucinated steps CTRLS enhances symbolic reasoning by reducing common algebraic mistakes—such as misapplied formulas and incorrect substitutions—as well as suppressing spurious steps that lack logical grounding. It preserves variable relationships across steps, avoids unnecessary formalisms, and maintains alignment with the problem's structure. Representative examples are provided in the original paper.

LLM-as-a-judge evaluation. To complement accuracy-based metrics, we additionally include a GPT-4 based rubric evaluation. For each configuration, we randomly sample 100 model outputs on GSM8K and let GPT-4 score them on a 1–10 scale across five criteria: self-reflection, algebra correctness, logical coherence, reduction of hallucinated steps, and overall quality. Table 7.3 summarises the averaged results for Qwen; detailed prompts and evaluation pipeline are deferred to the original paper.

7.6 Chapter Summary

We presented CTRLS, a principled framework for transition-aware chain-of-thought reasoning that casts step-wise generation as a latent-state Markov decision process. By modelling reasoning dynamics through distributional policies and optimizing latent transitions via reinforcement learning, CTRLS enables structured, interpretable, and controllable CoT generation. Our experiments demonstrate that CTRLS consistently improves reasoning accuracy, exploration efficiency, and robustness across math benchmarks, and further showcases explainability. Beyond performance, qualitative analyses highlight its ability to recover from symbolic errors, suppress spurious reasoning, and engage in self-reflective correction. We believe CTRLS offers a founda-

tion for more systematic and verifiable reasoning in large language models.

This chapter, in full, is a reprint of the material as it appears in "CTRLS: Chain-of-Thought Reasoning via Latent State Transition" by Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley, published in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026. The dissertation author was the primary investigator and author of this paper.

Chapter 8

Conclusion and Future Work

8.1 Summary of Contributions

This dissertation has presented six contributions across three thrusts, *representation* for multimodal understanding, *preference* for multimodal alignment, and *inference* for multimodal reasoning, unified by a single goal of making multimodal large language models more transparent, controllable, and trustworthy. Each thrust was addressed by two methodologically complementary chapters, and across all six a consistent argument emerges. The gains that matter for multimodal language models come not from scaling a fixed recipe, but from re-examining how signal enters the model, how it is supervised, and how it is searched at inference time.

Representation: modality-aware tuning is a prerequisite, not a refinement. CoMMIT (Chapter 2) and MDGD (Chapter 3) show that vanilla instruction tuning silently degrades the very visual capabilities that motivate MLLM design. CoMMIT attacks the problem at training time, measuring the relative learning of the feature encoder and the backbone language model with a multimodal balance coefficient and re-coordinating their learning rates when one side dominates. MDGD attacks it at the level of representation geometry, decoupling gradient updates by modality so that the effective rank of the pretrained visual subspace is preserved rather than collapsed. The information-theoretic preliminaries of Chapter 3 make explicit why visual forgetting happens, supervisory pressure flows through the textual loss while visual information faces a one-way com-

pression, and why coordinated, modality-decoupled updates can recover the lost capacity. Together the two chapters convert a previously implicit failure mode into something that can be measured, explained, and corrected.

Preference: alignment can move offline without giving up signal granularity. OCEAN (Chapter 4) and IRPO (Chapter 5) both operate in regimes where direct human feedback at scale is impractical. OCEAN reformulates chain-of-thought alignment as a Markov decision process over knowledge-graph states and derives an unbiased KG-IPS estimator with an accompanying variance bound, replacing freshly-collected human pairs with a structured external reward. IRPO enriches the *shape* of the signal rather than its source, extending pairwise preference optimization to in-context ranked lists with positional weighting and establishing its connection to importance sampling with reduced-variance, unbiased gradient estimates. The common lesson is that the bottleneck for alignment is rarely the optimization objective; it is the source and granularity of the reward, and that signal can come from places other than freshly-collected human pairs.

Inference: reasoning is a structured search problem, not a left-to-right autoregression. DeCoT (Chapter 6) treats spurious reasoning as a causal artifact, using a structural causal model and front-door adjustment to prune confounded paths without retraining. CTRLS (Chapter 7) casts chain-of-thought as transitions in a latent state space and explores plausible trajectories with distributional reinforcement learning. Both reject the assumption that "thinking step by step" yields a single privileged trajectory, and both show that modeling the structure of the reasoning space, rather than intervening around an opaque autoregression, improves accuracy, controllability, and interpretability at once.

Collectively, these contributions argue for a view of multimodal language models in which representation, alignment, and reasoning are not independent engineering stages but a single design surface: the way non-textual signal is encoded constrains what can be aligned, and the way preference is supplied constrains what can be reasoned over. The remainder of this chapter sketches how the tools developed here open onto a broader research program, much of which is already underway.

8.2 Future Directions

From a diagnostic to a constructive view. The information-bottleneck analysis underlying MDGD is, in this dissertation, primarily a *diagnostic*: it explains why visual capacity collapses but stops short of prescribing architectures or curricula that guarantee it will not. A natural next step is to make the lens constructive. Recent work treats the traceability and explainability of multimodal models as an information-theoretic object in its own right (Huang et al., 2025d), and steers modality preference at inference time through functional-entropy signals (Huang et al., 2026b). Turning these measurements into training-time objectives, supervising per-modality information flow directly, rather than diagnosing it after the fact, would close the loop between the analysis of Chapter 3 and the design of the next generation of encoders. The multimodal preference-optimization variants this program has begun to explore (Li et al., 2025c; Huang et al., 2025e) suggest that such objectives interact productively with alignment as well.

Offline and pluralistic alignment beyond curated structure. OCEAN draws its reward from a knowledge graph, which limits its reach to domains where curated structure already exists. The broader question is how far offline alignment can travel as that structure becomes weaker or more diverse. Pluralistic off-policy evaluation and alignment (Huang et al., 2025c) extends the offline-estimator idea to populations of heterogeneous preferences; active selection of preference queries (Kveton et al., 2025) attacks the granularity problem from the data side; and diffusion- and listwise-based preference objectives (Huang et al., 2026a) continue IRPO's move from pairwise bits toward richer feedback shapes. A unifying goal is an offline-alignment toolkit whose reward can come from knowledge graphs, simulator rollouts, program-execution traces, or population-level preference distributions interchangeably, each backed by estimator-level guarantees of the kind proved for KG-IPS in Chapter 4.

Latent-state reasoning and inference-time compute. CTRLS shows that treating reasoning as a latent-state Markov decision process makes exploration and control explicit, and DeCoT shows that causal structure can be imposed on a reasoning trace without retraining. The clearest

extension is to connect this latent-state view to the rapidly growing literature on inference-time compute. An empirical analysis of state-aware reasoning dynamics (Yu et al., 2025) indicates that the latent transitions CTRLS optimizes are visible and manipulable in existing models; online, inference-time action adaptation for agentic decision-making (Yu et al., 2026) and weakly-supervised, rollout-efficient policy optimization (Mundada et al., 2026) suggest that the latent-MDP formulation can be scaled to long-horizon, tool-using settings. A principled account of *how much* latent-state search to spend per problem, an inference-time analogue of the scaling laws that govern next-token pretraining, remains open, and is in our view the single most promising direction to come out of this thrust.

Toward agentic, multimodally-grounded systems. Finally, the three thrusts increasingly need to operate together inside a single agent that perceives, aligns, and reasons in the loop. Grounding multimodal reasoning in explicit structure, for example, aligning a model's reasoning to the scene graph of a complex visual input (Wang et al., 2026a), combines the representation and reasoning agendas, while agentic deployment exposes alignment to feedback that is sparse, delayed, and interactive rather than offline (Xia et al., 2025b; Wang et al., 2025b, 2026b). Extending the methods of this dissertation from vision-and-text to richer modalities (audio, structured documents, and embodied perception) and from single-turn benchmarks to persistent, tool-using agents is the longer-horizon program these contributions point toward. The recurring thesis, that the source of the signal, the form of its supervision, and the structure of the search matter more than scale alone, offers, we hope, a durable lens for that program.

REFERENCES

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE transactions on pattern analysis and machine intelligence*, 40 (12):2897–2905, 2018.

Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. A general framework for counterfactual learning-to-rank, 2019. URL `https://arxiv.org/abs/1805.00065`.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv preprint arXiv:2010.12688*, 2020.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. 2017. URL `https://openreview.net/forum?id=HyxQzBceg`.

Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13 (1):7240, 2023.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. *arXiv preprint arXiv:2306.04136*, 2023.

Chenjia Bai, Yang Zhang, Shuang Qiu, Qiaosheng Zhang, Kang Xu, and Xuelong Li. Online preference alignment for language models via count-based exploration. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a. URL https://arxiv.org/pdf/2204.05862.pdf.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b. URL https://arxiv.org/pdf/2212.08073.pdf.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

Aniruddha Bhargava, Lalit Jain, Branislav Kveton, Ge Liu, and Subhojyoti Mukherjee. Off-policy evaluation from logged human feedback. *arXiv preprint arXiv:2406.10030*, 2024.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

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 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. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Diego Carraro and Derek Bridge. Enhancing recommendation diversity by re-ranking with large language models. *ACM Trans. Recomm. Syst.*, October 2024. doi: 10.1145/3700604. URL https://doi.org/10.1145/3700604. Just Accepted.

Wen-Shuo Chao, Zhi Zheng, Hengshu Zhu, and Hao Liu. Make large language model a better ranker. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 918–929, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.51. URL https://aclanthology.org/2024.findings-emnlp.51/.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. *arXiv preprint arXiv:2311.11860*, 2023a.

Mingda Chen, Xilun Chen, and Wen-tau Yih. Efficient open domain multi-hop question answering with few-shot data synthesis. *arXiv preprint arXiv:2305.13691*, 2023b.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *arXiv preprint arXiv:2310.12086*, 2023c.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. *arXiv preprint arXiv:2406.09215*, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024b.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International conference on machine learning*, pages 834–843. PMLR, 2017.

Sanjiban Choudhury. Process reward models for llm agents: Practical framework and directions. *arXiv preprint arXiv:2502.10325*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023a.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2023b.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2023c.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. pages 2924–2936, June 2019. doi: 10.18653/v1/N19-1300. URL `https://aclanthology.org/N19-1300/`.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, et al. Language modeling is compression. *arXiv preprint arXiv:2309.10668*, 2023.

Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul G Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. *arXiv preprint arXiv:2407.07018*, 2024.

Xin Dong, Ruize Wu, Chao Xiong, Hai Li, Lei Cheng, Yong He, Shiyou Qian, Jian Cao, and Linjian Mo. Gdod: Effective gradient descent using orthogonal decomposition for multi-task learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 386–395, 2022.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

Muhammad Fawi. Curlora: Stable llm continual fine-tuning and catastrophic forgetting mitigation. *arXiv preprint arXiv:2408.14572*, 2024.

Chao Feng, Xinyu Zhang, and Zichu Fei. Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs. *arXiv preprint arXiv:2309.03118*, 2023.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance. *arXiv preprint arXiv:2305.17306*, 2023.

Zeyu Gan, Yun Liao, and Yong Liu. Rethinking external slow-thinking: From snowball errors to probability of correct reasoning, 2025. URL `https://arxiv.org/abs/2501.15602`.

Chongming Gao, Mengyao Gao, Chenxiao Fan, Shuai Yuan, Wentao Shi, and Xiangnan He. Process-supervised llm recommenders via flow-guided tuning, 2025. URL `https://arxiv.org/abs/2503.07377`.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Qitong Gao, Ge Gao, Juncheng Dong, Vahid Tarokh, Min Chi, and Miroslav Pajic. Off-policy evaluation for human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*, 2023.

Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint*

arXiv:2310.00603, 2023.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 198–206, 2018.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. INSPIRED: Toward sociable recommendation dialog systems. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.654. URL `https://aclanthology.org/2020.emnlp-main.654/`.

Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022.

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 720–730, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn

Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. Ralle: A framework for developing and evaluating retrieval-augmented large language models. *arXiv preprint arXiv:2308.10633*, 2023.

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng, Antoine Bosselut, and Mrinmaya Sachan. Towards a mechanistic interpretation of multi-step reasoning capabilities of language models, 2023. URL https://arxiv.org/abs/2310.14491.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models.

Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Songwen Hu, Ryan A. Rossi, Tong Yu, Junda Wu, Handong Zhao, Sungchul Kim, and Shuai Li. Interactive visualization recommendation with hier-sucb. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 313–321, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712746. doi: 10.1145/3696410.3714697. URL https://doi.org/10.1145/3696410.3714697.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. *arXiv preprint arXiv:2311.18445*, 2023.

Chengkai Huang, Hongtao Huang, Tong Yu, Kaige Xie, Junda Wu, Shuai Zhang, Julian Mcauley, Dietmar Jannach, and Lina Yao. A survey of foundation model-powered recommender systems: From feature-based, generative to agentic paradigms. *arXiv preprint arXiv:2504.16420*, 2025a.

Chengkai Huang, Junda Wu, Yu Xia, Zixu Yu, Ruhan Wang, Tong Yu, Ruiyi Zhang, Ryan A Rossi, Branislav Kveton, Dongruo Zhou, et al. Towards agentic recommender systems in the era of multimodal large language models. *arXiv preprint arXiv:2503.16734*, 2025b.

Chengkai Huang, Junda Wu, Zhouhang Xie, Yu Xia, Rui Wang, Tong Yu, Subrata Mitra, Julian McAuley, and Lina Yao. Pluralistic off-policy evaluation and alignment. *arXiv preprint arXiv:2509.19333*, 2025c.

Hongtao Huang, Chengkai Huang, Junda Wu, Tong Yu, Julian McAuley, and Lina Yao. Listwise preference diffusion optimization for user behavior trajectories prediction. *Advances in Neural Information Processing Systems*, 38:159383–159408, 2026a.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804, 2024a.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multimodal large language models. *arXiv preprint arXiv:2402.14683*, 2024b.

Zihan Huang, Junda Wu, Rohan Surana, Raghav Jain, Tong Yu, Raghavendra Addanki, David Arbour, Sungchul Kim, and Julian McAuley. Traceable and explainable multimodal large language models: An information-theoretic view. In *Second Conference on Language Modeling*, 2025d. URL https://openreview.net/forum?id=pQm66IPmeE.

Zihan Huang, Junda Wu, Rohan Surana, Tong Yu, David Arbour, Ritwik Sinha, and Julian McAuley. Image difference captioning via adversarial preference optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33746–33758, 2025e.

Zihan Huang, Xintong Li, Junda Wu, Rohan Surana, Tong Yu, Rui Wang, Julian McAuley, and Jingbo Shang. AMPS: Adaptive modality preference steering via functional entropy. 2026b. URL https://openreview.net/forum?id=XekKGk5Fcd.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Exploring the benefits of training expert language models over instruction tuning. *arXiv preprint arXiv:2302.03202*, 2023.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

Olivier Jeunen. Revisiting offline evaluation for implicit-feedback recommender systems. In *Proceedings of the 13th ACM conference on recommender systems*, pages 596–600, 2019.

Olivier Jeunen, Ivan Potapov, and Aleksei Ustimenko. On (normalised) discounted cumulative gain as an off-policy evaluation metric for top-n recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 1222–1233, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671687. URL https://doi.org/10.1145/3637528.3671687.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Chumeng Jiang, Jiayin Wang, Weizhi Ma, Charles L. A. Clarke, Shuai Wang, Chuhan Wu, and Min Zhang. Beyond utility: Evaluating llm as recommender, 2024a. URL https://arxiv.org/abs/2411.00331.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.

Wenyuan Jiang, Wenwei Wu, Le Zhang, Zixuan Yuan, Jian Xiang, Jingbo Zhou, and Hui Xiong. Killing two birds with one stone: Cross-modal reinforced prompting for graph and language tasks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1301–1312, 2024b.

Zhiying Jiang, Raphael Tang, Ji Xin, and Jimmy Lin. Inserting information bottlenecks for attribution in transformers. *arXiv preprint arXiv:2012.13838*, 2020.

Congyun Jin, Ming Zhang, Weixiao Ma, Yujiao Li, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, et al. Rjua-meddqa: A multimodal benchmark for medical document question answering and clinical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5218–5229, 2024.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. pages 2567–2577, November 2019. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 781–789, 2017.

Nitish Joshi, Koushik Kalyanaraman, Zhiting Hu, Kumar Chellapilla, He He, and Erran Li. Improving multi-hop reasoning in llms by learning from rich human feedback. 2024.

Yuta Kawakami, Manabu Kuroki, and Jin Tian. Instrumental variable estimation of average partial

causal effects. In *International Conference on Machine Learning*, pages 16097–16130. PMLR, 2023.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*, 2022.

Niki Kilbertus, Matt J Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33:20108–20119, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. URL `http://arxiv.org/abs/1412.6980`.

Anastasiia Klimashevskaia, Dietmar Jannach, Mehdi Elahi, and Christoph Trattner. A survey on popularity bias in recommender systems. *User Modeling and User-Adapted Interaction*, 34(5): 1777–1834, July 2024. ISSN 0924-1868. doi: 10.1007/s11257-024-09406-0. URL `https://doi.org/10.1007/s11257-024-09406-0`.

Dohwan Ko, Ji Soo Lee, Wooyoung Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. *arXiv preprint arXiv:2310.15747*, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Branislav Kveton, Xintong Li, Julian McAuley, Ryan Rossi, Jingbo Shang, Junda Wu, and Tong Yu. Active learning for direct preference optimization. *arXiv preprint arXiv:2503.01076*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023. URL `https://arxiv.org/pdf/2309.00267.pdf`.

Kuang-Huei Lee, Anurag Arnab, Sergio Guadarrama, John Canny, and Ian Fischer. Compressive visual representations. *Advances in Neural Information Processing Systems*, 34:19538–19552, 2021.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Chen Li, Yixiao Ge, Dian Li, and Ying Shan. Vision-language instruction tuning: A review and analysis. *Transactions on Machine Learning Research*, 2024a.

Jiachun Li, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. Focus on your question! interpreting and mitigating toxic cot problems in commonsense reasoning. *arXiv preprint arXiv:2402.18344*, 2024b.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. pages 19730–19742, 2023b.

Li Li, Peilin Cai, Ryan A. Rossi, Franck Dernoncourt, Branislav Kveton, Junda Wu, Tong Yu, Linxin Song, Tiankai Yang, Yuehan Qin, Nesreen K. Ahmed, Samyadeep Basu, Subhojyoti Mukherjee, Ruiyi Zhang, Zhengmian Hu, Bo Ni, Yuxiao Zhou, Zichao Wang, Yue Huang, Yu Wang, Xiangliang Zhang, Philip S. Yu, Xiyang Hu, and Yue Zhao. A personalized conversational benchmark: Towards simulating personalized conversations, 2025a. URL `https://openreview.net/forum?id=Te9wGKtakV`.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

Qintong Li, Zhiyong Wu, Lingpeng Kong, and Wei Bi. Explanation regeneration via information bottleneck. *arXiv preprint arXiv:2212.09603*, 2022.

Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9748–9758, Red Hook, NY, USA, 2018. Curran Associates Inc.

Ruosen Li and Xinya Du. Leveraging structured information for explainable multi-hop question answering and reasoning. *arXiv preprint arXiv:2311.03734*, 2023.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. From matching to generation: A survey on generative information retrieval, 2025b. URL https://arxiv.org/abs/2404.14851.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq Joty, and Soujanya Poria. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *arXiv preprint arXiv:2305.13269*, 2023c.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations*, 2024c. URL https://openreview.net/forum?id=cPgh4gWZlz.

Xintong Li, Chuhan Wang, Junda Wu, Rohan Surana, Tong Yu, Julian McAuley, and Jingbo Shang. Importance sampling for multi-negative multimodal direct preference optimization. *arXiv preprint arXiv:2509.25717*, 2025c.

Xintong Li, Junda Wu, Tong Yu, Rui Wang, Yu Wang, Xiang Chen, Jiuxiang Gu, Lina Yao, Julian McAuley, and Jingbo Shang. Commit: Coordinated multimodal instruction tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11533–11547, 2025d.

Zhifeng Li, Bowei Zou, Yifan Fan, and Yu Hong. Ufo: Unified fact obtaining for commonsense question answering. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023d.

Zhoubo Li, Ningyu Zhang, Yunzhi Yao, Mengru Wang, Xi Chen, and Huajun Chen. Unveiling the pitfalls of knowledge editing for large language models. *arXiv preprint arXiv:2310.02129*, 2023e.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Allen Lin, Jianling Wang, Ziwei Zhu, and James Caverlee. Quantifying and mitigating popularity bias in conversational recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1238–1247, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392365. doi: 10.1145/3511808.3557423. URL https://doi.org/10.1145/3511808.3557423.

Chu-Cheng Lin, Aaron Jaech, Xin Li, Matthew R Gormley, and Jason Eisner. Limitations of au-

toregressive models and their alternatives. *arXiv preprint arXiv:2010.11939*, 2020.

Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. 2024:19138–19162, 2024.

Xi Victoria Lin, Richard Socher, and Caiming Xiong. Multi-hop knowledge graph reasoning with reward shaping. *arXiv preprint arXiv:1808.10568*, 2018.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 36407–36433. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/72393bd47a35f5b3bee4c609e7bba733-Paper-Conference.pdf`.

Zhenqing Ling, Daoyuan Chen, Liuyi Yao, Yaliang Li, and Ying Shen. On the convergence of zeroth-order federated tuning for large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1827–1838, 2024.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Jialin Liu, Jianhua Wu, Jie Liu, and Yutai Duan. Learning attentional mixture of loras for language model continual learning. *arXiv preprint arXiv:2409.19611*, 2024a.

Jiongnan Liu, Jiajie Jin, Zihan Wang, Jiehan Cheng, Zhicheng Dou, and Ji-Rong Wen. Reta-llm: A retrieval-augmented large language model toolkit. *arXiv preprint arXiv:2306.05212*, 2023b.

Qi Liu, Bo Wang, Nan Wang, and Jiaxin Mao. Leveraging passage embeddings for efficient listwise reranking with large language models. In *THE WEB CONFERENCE 2025*, 2024b.

Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6566–6576, 2024c.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*, 2023c.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024d.

Tianyi Liu, Zuxuan Wu, Wenhan Xiong, Jingjing Chen, and Yu-Gang Jiang. Unified multimodal pre-training and prompt-based tuning for vision-language understanding and generation. *arXiv preprint arXiv:2112.05587*, 2021a.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021b.

Xu Liu, Tong Yu, Kaige Xie, Junda Wu, and Shuai Li. Interact with the explanations: Causal debiased explainable recommendation system. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 472–481, 2024e. URL `https://dl.acm.org/doi/pdf/10.1145/3616855.3635855`.

Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Information constraints on auto-encoding variational bayes. *Advances in neural information processing systems*, 31, 2018.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. 2021. URL `https://openreview.net/forum?id=uXa9oBDZ9V1`.

Yadong Lu, Chunyuan Li, Haotian Liu, Jianwei Yang, Jianfeng Gao, and Yelong Shen. An empirical study of scaling instruct-tuned large multimodal models. *arXiv preprint arXiv:2309.09958*, 2023.

James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. Aggregated momentum: Stability through passive damping. In *International Conference on Learning Representations*, 2018.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Sichun Luo, Bowei He, Haohan Zhao, Wei Shao, Yanlin Qi, Yinya Huang, Aojun Zhou, Yuxuan Yao, Zongpeng Li, Yuanzhang Xiao, et al. Recranker: Instruction tuning large language model as ranker for top-k recommendation. *ACM Transactions on Information Systems*, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Kaixin Ma, Hao Cheng, Xiaodong Liu, Eric Nyberg, and Jianfeng Gao. Open-domain question answering via chain of reasoning over heterogeneous knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5360–5374, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.

findings-emnlp.392. URL https://aclanthology.org/2022.findings-emnlp.392.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25: 4121–4134, 2022.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, 2023.

Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73, 2019.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. pages 9004–9017, 2023.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bodganov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. The song describer dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*, 2023.

Emily McMilin. Selection bias induced spurious correlations in large language models. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. URL https://openreview.net/forum?id=8nnaDv9dUli.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. pages 2381–2391, October-November 2018. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

IM Morato and A Mesaros. Macs-multi-annotator captioned soundscapes, 2021.

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

Gagan Mundada, Zihan Huang, Rohan Surana, Sheldon Yu, Jennifer Yuntong Zhang, Xintong Li, Tong Yu, Lina Yao, Jingbo Shang, Julian McAuley, et al. Ws-grpo: Weakly-supervised group-relative policy optimization for rollout-efficient reasoning. *arXiv preprint arXiv:2602.17025*, 2026.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Minh-Vuong Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. *arXiv preprint arXiv:2402.11199*, 2024.

Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Junyu Liu, and Benji Peng. From text to multimodality: Exploring the evolution and impact of large language models in medical practice. *arXiv preprint arXiv:2410.01812*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*, 2023.

Judea Pearl. Causal inference in statistics: An overview. 2009.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars

Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, 2021.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. Rankvicuna: Zero-shot listwise document reranking with open-source large language models, 2023. URL `https://arxiv.org/abs/2309.15088`.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus. *arXiv preprint arXiv:2304.04358*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Gowtham Ramesh, Makesh Sreedhar, and Junjie Hu. Single sequence prediction over reasoning graphs for multi-hop qa. *arXiv preprint arXiv:2307.00335*, 2023.

Vipula Rawte, Anku Rani, Harshad Sharma, Neeraj Anand, Krishnav Rajbangshi, Amit Sheth, and Amitava Das. Visual hallucination: Definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2403.17306*, 2024.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In

2007 15th European signal processing conference, pages 606–610. IEEE, 2007.

Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Mikhail Salnikov, Hai Le, Prateek Rajput, Irina Nikishina, Pavel Braslavski, Valentin Malykh, and Alexander Panchenko. Large language models meet knowledge graphs to answer factoid questions. *arXiv preprint arXiv:2310.02166*, 2023.

Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. Improving multi-hop question answering over knowledge graphs using knowledge base embeddings. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4498–4507, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL `https://arxiv.org/pdf/1707.06347.pdf`.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

Dominic Seyler, Mohamed Yahya, and Klaus Berberich. Knowledge questions from knowledge graphs. In *Proceedings of the ACM SIGIR international conference on theory of information retrieval*, pages 11–18, 2017.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Yibin Lei, Tianyi Zhou, Michael Blumenstein, and Daxin Jiang. Retrieval-augmented retrieval: Large language models are strong zero-shot retriever. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15933–15946, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.943. URL `https://aclanthology.org/2024.findings-acl.943/`.

Ashish Shenoy, Yichao Lu, Srihari Jayakumar, Debojeet Chatterjee, Mohsen Moslehpour, Pierce Chuang, Abhay Harpale, Vikas Bhardwaj, Di Xu, Shicong Zhao, et al. Lumos: Empowering multimodal llms with scene text recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5690–5700, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023a.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*, 2024.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*, 2023b.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Xin Su, Tiep Le, Steven Bethard, and Phillip Howard. Semi-structured chain-of-thought: Integrating multiple sources of knowledge for improved language model reasoning. *arXiv preprint arXiv:2311.08505*, 2023.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language

model with knowledge graph. *arXiv preprint arXiv:2307.07697*, 2023a.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*, 2023b.

Zhiqing Sun, Yikang Shen, Hongxin Zhang, Qinhong Zhou, Zhenfang Chen, David Daniel Cox, Yiming Yang, and Chuang Gan. SALMON: Self-alignment with principle-following reward models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=xJbsmB8UMx.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36, 2024b.

Rohan Surana, Junda Wu, Zhouhang Xie, Yu Xia, Harald Steck, Dawen Liang, Nathan Kallus, and Julian McAuley. From reviews to dialogues: Active synthesis for zero-shot llm-based conversational recommender system. *arXiv preprint arXiv:2504.15476*, 2025.

Richard S Sutton. Reinforcement learning: An introduction. *A Bradford Book*, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. pages 641–651, June 2018. doi: 10.18653/v1/N18-1059. URL https://aclanthology.org/N18-1059/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=qF7FlUT5dxa.

Hidenori Tanaka and Daniel Kunin. Noether's learning dynamics: Role of symmetry breaking in neural networks. *Advances in Neural Information Processing Systems*, 34:25646–25660, 2021.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023a.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500, 2024.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*, 2023b.

Gemma Team. Gemma. 2024a. doi: 10.34740/KAGGLE/M/3301. URL `https://www.kaggle.com/m/3301`.

Qwen Team. Qwen2.5: A party of foundation models, September 2024b. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. pages 809–819, June 2018. doi: 10.18653/v1/N18-1074. URL `https://aclanthology.org/N18-1074/`.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. pages 9568–9578, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*, 2022a.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022b.

Yu-Hsiang Tseng, Pin-Er Chen, Da-Chen Lian, and Shu-Kai Hsieh. The semantic relations in llms: An information-theoretic compression approach. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge)@ LREC-COLING-2024*, pages 8–21, 2024.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. *arXiv preprint arXiv:1905.07374*, 2019.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9073–9080, 2020.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023a.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc., 2023b. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf`.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. Learning to rank by optimizing ndcg measure. *Advances in neural information processing systems*, 22, 2009.

Chandra Shekhara Kaushik Valmeekam, Krishna Narayanan, Dileep Kalathil, Jean-Francois Chamberland, and Srinivas Shakkottai. Llmzip: Lossless text compression using large language models. *arXiv preprint arXiv:2306.04050*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Chuhan Wang, Xintong Li, Jennifer Yuntong Zhang, Junda Wu, Chengkai Huang, Lina Yao, Julian McAuley, and Jingbo Shang. Scenealign: Aligning multimodal reasoning to scene graphs in complex visual scenes. *arXiv preprint arXiv:2601.05600*, 2026a.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. Survey on factuality in large language models:

Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*, 2023a.

Dongsheng Wang, Miaoge Li, Xinyang Liu, MingSheng Xu, Bo Chen, and Hanwang Zhang. Tuning multi-mode token-level prompt alignment across modalities. *Advances in Neural Information Processing Systems*, 36, 2024a.

Fei Wang, Wenjie Mo, Yiwei Wang, Wenxuan Zhou, and Muhao Chen. A causal view of entity bias in (large) language models. *arXiv preprint arXiv:2305.14695*, 2023b.

Jianing Wang, Wenkang Huang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Xiang Li, and Ming Gao. Knowledge prompting in pre-trained language model for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3164–3177. Association for Computational Linguistics, 2022. URL https://doi.org/10.18653/v1/2022.emnlp-main.207.

Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. Boosting language models reasoning with chain-of-knowledge prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4958–4981. Association for Computational Linguistics, 2024b. URL https://doi.org/10.18653/v1/2024.acl-long.271.

Jianing Wang, Junda Wu, Yupeng Hou, Yao Liu, Ming Gao, and Julian McAuley. Instructgraph: Boosting large language models via graph-centric instruction tuning and preference alignment. pages 13492–13510, 2024c.

Jinyuan Wang, Junlong Li, and Hai Zhao. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. *arXiv preprint arXiv:2310.13552*, 2023c.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*, 2023d.

Rui Wang, Junda Wu, Yu Xia, Tong Yu, Ruiyi Zhang, Ryan Rossi, Subrata Mitra, Lina Yao, and Julian McAuley. Cacheprune: Teaching llms what not to follow via kv-cache editing, 2026b. URL https://arxiv.org/abs/2504.21228.

Ruoyu Wang, Junda Wu, Yu Xia, Tong Yu, Ryan A. Rossi, Julian McAuley, and Lina Yao. Dice: Dynamic in-context example selection in llm agents via efficient knowledge transfer, 2025a. URL https://arxiv.org/abs/2507.23554.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024d.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*, 2023e.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*, 2023f.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214, 2024e.

Yu Wang, Xinshuang Liu, Xiusi Chen, Sean O'Brien, Junda Wu, and Julian McAuley. Self-updatable large language models by integrating context into model parameters. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL `https://openreview.net/forum?id=aCPFCDL9QY`.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*, 2023g.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*, 2024f.

Fuchao Wei, Chenglong Bao, and Yang Liu. Stochastic anderson mixing for nonconvex stochastic optimization. *Advances in Neural Information Processing Systems*, 34:22995–23008, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=_VjQlMeSB_J`.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*, 2017.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

Yilin Wen, Zifeng Wang, and Jimeng Sun. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. *arXiv preprint arXiv:2308.09729*, 2023.

Jason Weston and Sainbayar Sukhbaatar. System 2 attention (is something you might need too). *ArXiv*, abs/2311.11829, 2023. URL `https://api.semanticscholar.org/CorpusID:265295357`.

Stephen J Wright. Coordinate descent algorithms. *Mathematical programming*, 151(1):3–34, 2015.

Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. Ctrls: Chain-of-thought reasoning via latent state-transition.

Junda Wu, Tong Yu, and Shuai Li. Deconfounded and explainable interactive vision-language retrieval of complex scenes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2103–2111, 2021a. URL `https://dl.acm.org/doi/pdf/10.1145/3474085.3475366`.

Junda Wu, Canzhe Zhao, Tong Yu, Jingyang Li, and Shuai Li. Clustering of conversational bandits for user preference learning and elicitation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2129–2139, 2021b.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448, 2022a.

Junda Wu, Zhihui Xie, Tong Yu, Handong Zhao, Ruiyi Zhang, and Shuai Li. Dynamics-aware adaptation for reinforcement learning based cross-domain interactive recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 290–300, 2022b.

Junda Wu, Rui Wang, Handong Zhao, Ruiyi Zhang, Chaochao Lu, Shuai Li, and Ricardo Henao. Few-shot composition learning for image retrieval with prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4729–4737, 2023a.

Junda Wu, Cheng-Chun Chang, Tong Yu, Zhankui He, Jianing Wang, Yupeng Hou, and Julian McAuley. Coral: collaborative retrieval-augmented large language models improve long-tail recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discov-*

ery and Data Mining, pages 3391–3401, 2024a.

Junda Wu, Hanjia Lyu, Yu Xia, Zhehao Zhang, Joe Barrow, Ishita Kumar, Mehrnoosh Mirtaheri, Hongjie Chen, Ryan A Rossi, Franck Dernoncourt, et al. Personalized multimodal large language models: A survey. *arXiv preprint arXiv:2412.02142*, 2024b.

Junda Wu, Zachary Novack, Amit Namburi, Jiaheng Dai, Hao-Wen Dong, Zhouhang Xie, Carol Chen, and Julian McAuley. Futga: Towards fine-grained music understanding through temporally-enhanced generative augmentation. pages 107–111, 2024c.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087, 2024d.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36, 2024e.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A Rossi, Ruiyi Zhang, et al. Visual prompting in multimodal large language models: A survey. *arXiv preprint arXiv:2409.15310*, 2024f.

Junda Wu, Jessica Echterhoff, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, and Julian McAuley. Pdb-eval: An evaluation of large multimodal models for description and explanation of personalized driving behavior. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 242–248. IEEE, 2025a.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, et al. Ocean: Offline chain-of-thought evaluation and alignment in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Junda Wu, Rohan Surana, Zhouhang Xie, Yiran Shen, Yu Xia, Tong Yu, Ryan A. Rossi, Prithviraj Ammanabrolu, and Julian McAuley. In-context ranking preference optimization. In *Second Conference on Language Modeling*, 2025c. URL `https://openreview.net/forum?id=L2NPhLAKEd`.

Junda Wu, Yu Xia, Tong Yu, Xiang Chen, Sai Sree Harsha, Akash V Maharaj, Ruiyi Zhang, Victor Bursztyn, Sungchul Kim, Ryan A Rossi, et al. Doc-react: Multi-page heterogeneous document question-answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 67–78, 2025d.

Junda Wu, Yuxin Xiong, Xintong Li, Yu Xia, Ruoyu Wang, Yu Wang, Tong Yu, Sungchul Kim, Ryan A Rossi, Lina Yao, et al. Mitigating visual knowledge forgetting in mllm instruction-tuning

via modality-decoupled gradient descent. *arXiv preprint arXiv:2502.11740*, 8, 2025e.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. β-dpo: Direct preference optimization with dynamic β. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024g.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation, 2024h. URL https://arxiv.org/abs/2305.19860.

Suhang Wu, Minlong Peng, Yue Chen, Jinsong Su, and Mingming Sun. Eva-kellm: A new benchmark for evaluating knowledge editing of llms. *arXiv preprint arXiv:2308.09954*, 2023b.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*, 2024i.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2023c. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b8c90b65739ae8417e61eadb521f63d5-Paper-Conference.pdf.

Yu Xia, Junda Wu, Tong Yu, Sungchul Kim, Ryan A. Rossi, and Shuai Li. User-regulation deconfounded conversational recommender system with bandit feedback. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 2694–2704, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599539. URL https://doi.org/10.1145/3580305.3599539.

Yu Xia, Rui Wang, Xu Liu, Mingyan Li, Tong Yu, Xiang Chen, Julian McAuley, and Shuai Li. Beyond chain-of-thought: A survey of chain-of-x paradigms for llms. *arXiv preprint arXiv:2404.15676*, 2024.

Yu Xia, Subhojyoti Mukherjee, Zhouhang Xie, Junda Wu, Xintong Li, Ryan Aponte, Hanjia Lyu, Joe Barrow, Hongjie Chen, Franck Dernoncourt, et al. From selection to generation: A survey of llm-based active learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14552–14569, 2025a.

Yu Xia, Yiran Jenny Shen, Junda Wu, Tong Yu, Sungchul Kim, Ryan A. Rossi, Lina Yao, and Julian McAuley. SAND: Boosting LLM agents with self-taught action deliberation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3062–3077, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.152. URL https://aclanthology.org/

2025.emnlp-main.152/.

Yu Xia, Junda Wu, Sungchul Kim, Tong Yu, Ryan A Rossi, Haoliang Wang, and Julian McAuley. Knowledge-aware query expansion with large language models for textual and relational retrieval, 2025c.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*, 2023.

Kaige Xie, Tong Yu, Haoliang Wang, Junda Wu, Handong Zhao, Ruiyi Zhang, Kanak Mahadik, Ani Nenkova, and Mark Riedl. Few-shot dialogue summarization via skeleton-assisted prompt transfer in prompt tuning. pages 2408–2421, 2024a.

Zhouhang Xie, Junda Wu, Hyunsik Jeon, Zhankui He, Harald Steck, Rahul Jha, Dawen Liang, Nathan Kallus, and Julian Mcauley. Neighborhood-based collaborative filtering for conversational recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 1045–1050, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688191. URL https://doi.org/10.1145/3640457.3688191.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-seng Chua. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *arXiv preprint arXiv:2304.14732*, 2023.

An Yan, Zhengyuan Yang, Junda Wu, Wanrong Zhu, Jianwei Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Julian McAuley, Jianfeng Gao, et al. List items one by one: A new data source and learning paradigm for multimodal llms. *arXiv preprint arXiv:2404.16375*, 2024.

Ting Yang and Li Chen. Unleashing the retrieval potential of large language models in conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, pages 43–52, 2024.

Zhen Yang, Yongbin Liu, and Chunping Ouyang. Causal interventions-based few-shot named entity recognition. *arXiv preprint arXiv:2305.01914*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380, 2018.

Zhou Yang, Zhengyu Qi, Zhaochun Ren, Zhikai Jia, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. Exploring information processing in large language models: Insights from information bottleneck theory. *arXiv preprint arXiv:2501.00999*, 2025.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. No preference left behind: Group distributional preference optimization. *arXiv preprint arXiv:2412.20299*, 2024a.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*, 2023a.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023b.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. 2023c. URL `https://openreview.net/forum?id=WE_vluYUL-X`.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024b. URL `https://arxiv.org/abs/2408.01800`.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Advances in Neural Information Processing Systems*, 36, 2024.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian McAuley. Explainable chain-of-thought reasoning: An empirical analysis on state-aware reasoning dynamics. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.904. URL `https://aclanthology.org/2025.findings-emnlp.904/`.

Sheldon Yu, Junda Wu, Xintong Li, Nikki Lijing Kuang, Sizhe Zhou, Tong Yu, Jiawei Han, Jingbo Shang, and Julian McAuley. Olivia: Online learning via inference-time action adaptation for decision making in llm react agents. 2026. URL `https://arxiv.org/abs/2605.11169`.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn.

Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*, 2023.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. Back to the future: Towards explainable temporal reasoning with large language models. *arXiv preprint arXiv:2310.01074*, 2023a.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=EdIGMCHk4l.

Junkun Yuan, Xu Ma, Ruoxuan Xiong, Mingming Gong, Xiangyu Liu, Fei Wu, Lanfen Lin, and Kun Kuang. Instrumental variable-driven domain generalization with unobserved confounders. *ACM Transactions on Knowledge Discovery from Data*, 2023c.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations. *arXiv preprint arXiv:2306.04618*, 2023d.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. Counterfactual generator: A weakly-supervised method for named entity recognition. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.590. URL https://aclanthology.org/2020.emnlp-main.590.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. Vpgtrans: Transfer visual prompt generator across llms. *Advances in Neural Information Processing Systems*, 36, 2024a.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh.

Improving the adversarial robustness of nlp models by information bottleneck. *arXiv preprint arXiv:2206.05511*, 2022.

Chen Zhang, Dading Chong, Feng Jiang, Chengguang Tang, Anningzhe Gao, Guohua Tang, and Haizhou Li. Aligning language models using follow-up likelihood as reward signal. *arXiv preprint arXiv:2409.13948*, 2024b.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Hongbo Zhang, Han Cui, Guangsheng Bao, Linyi Yang, Jun Wang, and Yue Zhang. Direct value optimization: Improving chain-of-thought reasoning in llms with refined values. *arXiv preprint arXiv:2502.13723*, 2025a.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*, 2023b.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4373–4382, 2020.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B Tenenbaum, and Chuang Gan. Planning with large language models for code generation. In *The Eleventh International Conference on Learning Representations*, 2023c.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. Mitigating language model hallucination with interactive question-knowledge alignment, 2023d.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28, 2015.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356, 2024c.

Yipeng Zhang, Xin Wang, Hong Chen, Jiapei Fan, Weigao Wen, Hui Xue, Hong Mei, and Wenwu Zhu. Large language model with curriculum reasoning for visual concept recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6269–6280, 2024d.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*, 2024e.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen,

and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space, 2025b. URL https://arxiv.org/abs/2505.15778.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*, 2023e.

Zongmeng Zhang, Yufeng Shi, Jinhua Zhu, Wengang Zhou, Xiang Qi, Houqiang Li, et al. Trustworthy alignment of retrieval-augmented large language models via reinforcement learning. In *Forty-first International Conference on Machine Learning*.

Hang Zhao, Yifei Xin, Zhesong Yu, Bilei Zhu, Lu Lu, and Zejun Ma. Slit: Boosting audio-text pre-training via multi-stage learning and instruction tuning. *arXiv preprint arXiv:2402.07485*, 2024a.

Ruochen Zhao, Shafiq Joty, Yongjie Wang, and Prathyusha Jwalapuram. Towards causal concepts for explaining language models, 2023a. URL https://openreview.net/forum?id=xYy2l4tiOe.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, 2023b.

Yang Zhao, Yixin Wang, and Mingzhang Yin. Ordinal preference optimization: Aligning human preferences via ndcg. *arXiv preprint arXiv:2410.04346*, 2024b.

Junhao Zheng, Qianli Ma, Zhen Liu, Binquan Wu, and Huawen Feng. Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer. *arXiv preprint arXiv:2401.09181*, 2024.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*, 2024a.

Jiacong Zhou, Xianyun Wang, and Jun Yu. Optimizing preference alignment with differentiable ndcg ranking. *arXiv preprint arXiv:2410.18127*, 2024b.

Banghua Zhu, Michael I Jordan, and Jiantao Jiao. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. *arXiv preprint arXiv:2401.16335*, 2024a.

Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. *arXiv preprint arXiv:2402.12048*, 2024b.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Shijie Zhu, Hui Zhao, Pengjie Wang, Hongbo Deng, Jian Xu, and Bo Zheng. Gradient deconfliction via orthogonal projections onto subspaces for multi-task learning. 2022a.

Wang Zhu, Jesse Thomason, and Robin Jia. Chain-of-questions training with latent answers for robust multistep question answering. *arXiv preprint arXiv:2305.14901*, 2023a.

Xianchao Zhu, Tianyi Huang, Ruiyuan Zhang, and William Zhu. Wdibs: Wasserstein deterministic information bottleneck for state abstraction to balance state-compression and performance. *Applied Intelligence*, pages 1–14, 2022b.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

Yin Zhu, Zhiling Luo, and Gong Cheng. Furthest reasoning with plan assessment: Stable reasoning path with retrieval-augmented large language models. *arXiv preprint arXiv:2309.12767*, 2023b.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. *arXiv preprint arXiv:2310.14122*, 2023.

Yuchen Zhuang, Xiang Chen, Tong Yu, Saayan Mitra, Victor Bursztyn, Ryan A Rossi, Somdeb Sarkhel, and Chao Zhang. Toolchain*: Efficient action space navigation in large language models with a* search. In *The Twelfth International Conference on Learning Representations*.