Towards Trustworthy and Reliable Language Models

Ruochen Zhao

College of Computing and Data Science

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

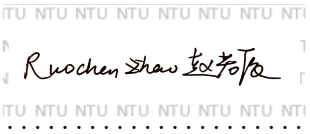
2025

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

01/08/2024

Date



Ruochen Zhao

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

01/08/2024

Date

Prof. Aixin Sun

Authorship Attribution Statement

This thesis contains material from 4 paper(s) published in the following papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as Ruochen Zhao, Shafiq Joty, Yongjie Wang, Tan Wang, Explaining Language Models' Predictions with High-Impact Concepts in Findings of the Association for Computational Linguistics: EACL 2024, pages 995–1012, online.

The contributions of the co-authors are as follows:

- I proposed the idea, designed the study, ran the experiments, and provided the draft paper.
- Prof. Joty had regular discussions with me and helped refine the idea, revise the paper, and supervise the entire project.
- Dr. Yongjie Wang helped refine the paper multiple times, proposing ablations and theories to prove the effectiveness of the results.
- Dr. Tan Wang had regular discussions with me to further refine the paper, providing structural changes and adding T5 and Llama models to further support the results.

Chapter 4 is published as Ruochen Zhao*, Xingxuan Li*, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. 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, Toronto, Canada. Association for Computational Linguistics.

The contributions of the co-authors are as follows:

- I proposed the initial idea and provided the draft paper.
- Xingxuan Li and I conducted experiments together, where he mainly worked on using Google to retrieve facts and incorporate into the framework.
- Xingxuan Li had regular discussions with me throughout the progress of the work, wrote sections of the draft paper, and helped me revise it.
- Prof. Shafiq Joty had regular discussions with me throughout the progress of the work and supervised the project.
- Chengwei Qin provided valuable feedback of the work.
- Dr. Lidong Bing gave very useful suggestions on the experiments and helped revise drafts of the paper.

Chapter 5 is published as Xingxuan Li*, Ruochen Zhao*, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, Lidong Bing, Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources, The Twelfth International Conference on Learning Representations (ICLR) 2024.

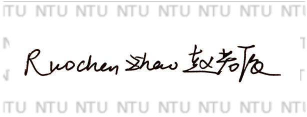
The contributions of the co-authors are as follows:

- Xingxuan Li proposed the initial idea and provided the draft paper.
- Xingxuan Li, Yew Ken Chia and I conducted experiments together. I focused on the fever experiments and human study, Yew Ken Chia focused on the table QA task, and Xingxuan did everything else, including HotpotQA and the medical and physical datasets.
- Xingxuan Li, Yew Ken Chia, Dr. Bosheng Ding, Dr. Lidong Bing, and I had regular discussions throughout the project and refined the paper together.
- Dr. Lidong Bing helped supervise the entire project, gave valuable suggestions, and helped revise the paper.
- Prof. Shafiq Joty and Prof. Soujanya Poria helped give valuable feedback to the paper.

Chapter 6 is published as Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving Multimodal Information for Augmented Generation: A Survey. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4736–4756, Singapore. Association for Computational Linguistics.

The contributions of the co-authors are as follows:

- I organized the project, wrote the introduction and conclusion, and refined the final paper version.
- Hailin Chen helped refine the initial survey scope and wrote the section on image.
- Weishi Wang and Xuan Long Do wrote the section on code.
- Fangkai Jiao wrote the section on video.
- Chengwei Qin wrote the section on future directions.
- Dr. Bosheng Ding and Xiaobao Guo wrote the background section.
- Minzhi Li wrote the section on speech.
- Xingxuan Li and I wrote the section on structured knowledge.
- Prof. Shafiq Joty gave valuable feedback to the survey and had regular meetings with us throughout the project.



01/08/2024 . . . .

Date

Ruochen Zhao

Acknowledgements

I wish to express my deepest gratitude to my supervisors Prof. Shafiq Joty Rayhan and Prof. Aixin Sun. Their guidance and patience provided valuable support to my academic and personal growth. I will cherish the research skills and scientific mindset they taught me. I am also grateful to AI Singapore for supporting my Ph.D. study.

At the same time, I would like to thank my mentors, Dr. Lidong Bing and Dr. Wenxuan Zhang, at Alibaba DAMO Academy during our collaborations. Their guidance was invaluable for my projects.

Moreover, I would like to thank my collaborators and labmates, Xingxuan Li, Mathieu Ravaut, Chengwei Qin, Hailin Chen, Tan Wang, Yew Ken Chia, Yongjie Wang, Bosheng Ding, and so on. They provided valuable suggestions and detailed help on my research journey, always generously helping me revise my draft papers and codes.

Last but not least, I would like to thank my family, partner Gongpu Zhang, best friend Deng Pan, Yitian Zhang, and my cat Bagel, and so many others for their emotional support, which was crucial throughout the journey.

Abstract

This thesis addresses the critical challenge of developing trustworthy and reliable Natural Language Processing (NLP) systems, specifically the newly emerged Large Language Models (LLMs). As LLMs become increasingly prevalent in various domains, the need for transparent, interpretable, and controllable AI systems has never been more pressing. However, the complexity of LLMs, the compositional nature of language, and the potential for hallucinations pose significant obstacles to achieving these goals.

To increase user trust of AI systems in real-life deployment, we hope to enhance the trustworthiness and reliability of LLMs without requiring model revisions or compromising performance. Motivated by this overarching goal, we delve into two main goals that enhance trustworthiness, providing user-friendly explanations of the LLM's decisions and controlling the LLM's behaviors. Specifically, we raise three main research questions: How can we disentangle the true reasons behind LLM decisions from the complex architecture and vast number of parameters? How can we provide user-friendly explanations for LLM generations? How can we increase LLM controllability with minimal interventions?

Motivated by these questions, we proposed several novel frameworks and conducted a comprehensive survey. Firstly, to provide users with more transparent insights into LLM decision-making processes, we introduce the High-Impact Concepts framework, which addresses the need for causal and interpretable explanations of LLM behaviors, moving beyond correlational explanations to provide more faithful interpretations. Then, besides simply explaining the LLM decisions, we further propose the Verify-and-Edit framework, which enhances LLM controllability using the explanations. To increase user confidence in LLM output, it tackles the challenge of factual correctness in LLM output for knowledge-intensive tasks, improving reliability without modifying the underlying model. To improve the Verify-and-Edit framework to include various formats of knowledge and further improve reliability, we propose the Chain of Knowledge (CoK) framework,

which focuses on reducing hallucinations and improving LLM controllability by dynamically incorporating grounding information from diverse sources. Finally, recognizing the growing importance of multimodal interactions, we present a comprehensive survey on retrieving multimodal information for augmented generation. This survey is motivated by the need to understand how diverse modalities can be used to increase LLM controllability and groundness, thus improving its robustness and reliability.

Experimental results demonstrate the effectiveness of these approaches in improving the interpretability, factual accuracy, and controllability of LLMs across various tasks and domains. This thesis hopes to not only advance our theoretical understanding of LLMs, but also offer practical tools to improve their trustworthiness and reliability. As AI systems continue to evolve and permeate various aspects of society, the methodologies and insights presented in this thesis pave the way for the responsible and effective deployment of LLMs, ensuring that these powerful technologies can be harnessed to their full potential while maintaining the trust and confidence of users and stakeholders.

Contents

Acknowledgements xi

Abstract xiii

List of Figures xix

List of Tables xxi

Symbols and Acronyms xxiii

1 Introduction 1
- 1.1 Motivations . . . 3
- 1.2 Major Contributions . . . 5
- 1.3 Outline of the Thesis . . . 7

2 Literature Review 9
- 2.1 Explaining LLM's decisions . . . 9
  - 2.1.1 General Explainability Methods . . . 10
    - 2.1.1.1 XAI methods for AI in General: . . . 10
      - Current Challenges: . . . 12
    - 2.1.1.2 XAI methods specialized in NLP: . . . 13
      - Current Challenges: . . . 14
  - 2.1.2 Causal explainability methods . . . 14
    - 2.1.2.1 Causal XAI methods for AI in General . . . 15
      - Current challenges: . . . 16
    - 2.1.2.2 Causal XAI methods specialized in NLP . . . 17
      - Current challenges: . . . 18
- 2.2 Making LLM Generations More Factual . . . 18
  - Reducing Hallucinations in Text Generations: . . . 18
  - Reducing Hallucinations in CoTs: . . . 20
  - Existing Challenges: . . . 21
- 2.3 Making LLM Generations More Controllable . . . 21
  - Controlling LM Outputs: . . . 21
  - Controlling LLM Outputs: . . . 22

Existing challenges: . . . 23

3 Explaining Language Model Predictions with High-Impact Concepts 25
3.1 Chapter Background . . . 26
3.2 Preliminaries . . . 28
3.2.1 Concept-based Explanations . . . 28
3.2.2 Concept Bottleneck Models . . . 29
3.3 Methodology . . . 30
3.3.1 Defining Impact . . . 30
3.3.2 Optimizing for Impact . . . 31
3.3.3 Visualizing Concepts via Impact . . . 33
3.3.4 Evaluating Impact of Concepts . . . 33
3.4 Experiment Setup . . . 34
3.4.1 Datasets and Models . . . 34
3.4.2 Metrics . . . 36
3.4.3 Baselines and Hyperparameters Used . . . 37
3.4.4 Training details . . . 39
3.4.5 Run-time . . . 39
3.5 Results and Analysis . . . 40
3.5.1 Sanity Check . . . 41
3.5.2 Quantitative Results on Text Classification . . . 44
3.5.3 Qualitative Analysis of Text Classification . . . 44
3.5.4 Generalization to Concept Insertion . . . 45
3.5.5 Human Study . . . 46
3.5.6 Ablation Study . . . 49
3.6 Conclusions . . . 54

4 Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework 55
4.1 Chapter Background . . . 55
4.2 Verify-and-Edit Framework . . . 58
4.2.1 Deciding when to edit . . . 59
4.2.2 How to edit a specific rationale . . . 60
4.2.3 Answering again . . . 61
4.3 Experiment Setup . . . 62
4.3.1 Reasoning tasks . . . 62
4.3.2 Compared methods . . . 63
4.4 Results and Analysis . . . 65
4.4.1 Using Self-Consistency: know when it doesn't know . . . 65
4.4.2 Results on HotpotQA . . . 66
4.4.3 Results on 2WikiMultiHop . . . 67
4.4.4 Results on fact verification . . . 68
4.4.5 Cost considerations . . . 69

4.4.6 Evaluating the reasoning chains with human study 70
4.4.7 Ablation study: editing at different consistency thresholds . 70
4.5 Conclusions 72
4.6 Prompts Used 73
4.6.1 HotpotQA 73
4.6.1.1 Few-shot prompt 73
4.6.1.2 CoT, CoT-SC prompt 74
4.6.1.3 Verifying Question Generation prompt 75
4.6.1.4 Verifying Answer Generation (Rationale Editing) prompt 75
4.6.2 2WikiMultihop 77
4.6.2.1 Few-shot prompt 77
4.6.3 CoT, CoT-SC prompt 77
4.6.3.1 Verifying Question Generation prompt 78
4.6.3.2 Verifying Answer Generation (Rationale Editing) prompt 79
4.6.4 Fever 79
4.6.4.1 Few-shot prompt 79
4.6.4.2 CoT, CoT-SC prompt 80
4.6.4.3 Verifying Question Generation prompt 81
4.6.4.4 Verifying Answer Generation (Rationale Editing) prompt 81

5 Chain-of-Knowledge: a Follow-up on Verify-and-Edit to Diverse Knowledge Souces 83
5.1 Chapter Background 83
5.2 Methodology 85
5.3 Experiments 86
5.3.1 Setup 86
5.3.2 Databases and AQGs 86
5.3.2.1 ScienceQA Physics 87
Instruction-tuning Dataset 87
Data Example in Instruction-tuning Dataset 87
Query Execution 87
5.3.2.2 PhysicsClassroom (Natural Sentence): 88
5.3.3 Experimental Results 88
5.3.4 Conclusion 89
5.3.5 Limitations 89

6 Retrieving Multimodal Information for Augmented Generation: A Survey 91
6.1 Chapter Introduction 91
6.2 Definitions and Background 93
6.2.1 Multimodal Learning 93

6.2.2 Retrieval-Augmented Generation (RAG) 94
6.2.3 Search Criteria and Results 94
6.3 Multimodal Retrieval-Augmented Generation 96
6.3.1 Image 96
6.3.2 Code 98
6.3.3 Structured Knowledge 99
6.3.4 Audio 101
6.3.5 Video 102
6.4 Future Directions 104
6.4.1 Retrieval Augmented Multimodal Reasoning 104
6.4.2 Building a Multimodal Knowledge Index 104
6.4.3 Pretraining with Multimodal Retrieval 105
6.5 Conclusions 106

7 Conclusions and Future Directions **107**
7.1 Conclusions 107
7.2 Future Directions 108

List of Author's Awards, Patents, and Publications **115**

Bibliography **117**

List of Figures

1.1 Overview of the Thesis. 7

3.1 Illustration of concept-based explanations that result in high impact (green line) or not (red line) when explaining the LLMs in a sentiment classification task. 26
3.2 The overall concept generation process of a concept bottleneck model. 29
3.3 Illustration of the causal graph indicating the confounding association in explanation models. Blue is a real-life example. Green is the correspondence in a movie review classification task. 30
3.4 Convolutional Neural Network used for classifying the toy dataset. . 42
3.5 Examples from the toy dataset and concepts discovered. 43
3.6 Qualitative comparison from AG-News: "World" news misclassified as "Sports" by BERT. 45
3.7 Human study instructions for plain examples. 47
3.8 Human study instructions for *HI-concept* augmented examples. . . . 47
3.9 Human study question and answer. 48
3.10 Human study question and answer. 48
3.11 Effects of concept insertion on accuracy on AG-News dataset. Each figure represents a different model where the number of inserted concepts (x-axis) is plotted against accuracy (y-axis). 49
3.12 Layer-wise effective number of concepts, RAcc ↑, I($\mathcal{C}$) ↑, and Δ Acc ↑. 51
3.13 Wordclouds of concepts generated on the 9th (left) and 12th (right) layer. The 9th layer includes a government concept, a China concept, and an Adjective (mostly) concept. The 12th layer includes a sports concept, a technology concept, and a political concept. 52
3.14 Layer-wise Topic Coherence Comparison. 52
3.15 Concept-wise effective number of concepts, RAcc ↑, I($\mathcal{C}$) ↑, and Δ Acc ↑. 53
3.16 Concept-wise Topic Coherence Comparison. 53

4.1 The Verify-and-Edit framework consists of five steps: (1) pass predictions with lower-than-average consistency to the next stages while leaving highly consistent predictions as-is; (2) produce verifying questions; (3) retrieve external knowledge; (4) edit rationales with informed answers; and (5) produce new predictions. 57

4.2 Kernal density estimation plots for consistency on the Adversarial **HotpotQA** dataset. With kernal estimation, the curve extends its true distribution's range, which is from 0 to 5 (as we sampled 5 paths). 66
4.3 Example Screenshot of Human Evaluation User Interface. 71
4.4 Ablation study on the effect of various consistency thresholds on task performances on Adversarial HotpotQA 72

5.1 An overview of the CoK framework with an example in Physics. . . 90

6.1 Paper trend analysis 95

List of Tables

3.1 A summary of the datasets. 34
3.2 Hyperparameters for finetuning BERT model. 35
3.3 A summary of runtime (in seconds) on datasets for BERT. 39
3.4 Faithfulness (Acc) and Causality (CACE, ΔAcc) evaluation on the toy dataset. Cls.Acc denotes the original classification model's accuracy. 40
3.5 Faithfulness (Acc, Precision, Recall, F1, Completeness) and causality (CACE, ΔAcc) evaluation of different text classification methods. The best result is bolded, and the second-best result is underlined. . 41
3.6 Generated concepts with Average Impact (CACE) from AG-News dataset, BERT model. CS is ConceptSHAP, HI-C is *HI-concept*. Each line is one concept, represented by keywords, which are ordered by descending importance. 43
3.7 Human study for explainability evaluation. 45
3.8 Ablation on BERT for IMDB with faithfulness (Acc) and impact (CACE, ΔAcc) evaluation. 50

4.1 Results on the Adversarial **HotpotQA** dataset. The best result for each model is underlined and the best result overall is bolded. ΔEM represents the improvement on Exact Match from the CoT-SC baseline. The top two rows uses the PaLM model and the rest uses the GPT-3 davinci-003 model. 67
4.2 Results on **2WikiMultiHopQA** dataset. ΔEM represents the improvement on Exact Match from the CoT-SC baseline. All experiment uses the GPT-3 davinci-003 model. 68
4.3 Results on **Fever** dataset. ΔAccuracy represents the improvement on Accuracy from the CoT-SC baseline. The top two rows use the PaLM model and the rest of the rows use the GPT-3 davinci-003 model. 68
4.4 Human study for factuality of CoTs on the HotpotQA dataset. "Ours" refers to the Verify-and-Edit model with Google retrieval. . 70

5.1 Main experimental results on MMLU Physics. Standard refers to standard few-shot prompting. CoT refers to Chain-of-thought prompting [1]. CoT-SC refers to CoT with self-consistency [2]. VE refers to Verify-and-Edit in Chapter 4. Acc.: accuracy. Δ Acc.: change in accuracy compared to CoT. 88

6.1 Paper statistics. Number in parenthesis is the number before manual filtering. "Google" represents searching on google scholar and manually filtering. "Total analyzed" represents the number of total papers after manual filtering 95

Symbols and Acronyms

Symbols

$\mathcal{R}^n$	the n-dimensional Euclidean space
$\|\cdot\|$	the 2-norm of a vector or matrix in Euclidean space
∇f	the gradient vector
$\mathbb{E}[\,\cdot\,]$	the expectation
$\|\cdot\|$	the absolute value

Acronyms

AI	Artificial Intelligence
CoT	Chain of Thought prompting
NLP	Natural Language Processing
LM	Language Models
LLM	Large Language Models
ML	Machine Learning
RAG	Retrieval Augmented Generation
i.i.d.	independent and identically distributed

Chapter 1

Introduction

The rapid advancement of Natural Language Processing (NLP) systems, particularly large language models (LLMs), has revolutionized various domains of artificial intelligence (AI). However, as these systems become increasingly sophisticated and widespread, the critical need for trustworthiness and reliability in AI has reached the forefront of research and public discourse. We define trustworthy and reliable AI systems as those that consistently perform as intended, adhere to ethical principles, and maintain transparency in their decision-making processes.

In this thesis, the term "Trustworthy" focuses on the end users, while the term "Reliable" focuses on the NLP systems' outputs. Specifically, we define a trustworthy NLP system to be a system that can be trusted by the end users. A typical user places more trust in systems whose decision-making mechanisms are understandable. Therefore, we could pin down trustworthiness to include two main aspects: "interpretablity" and "faithfulness". As mentioned in Miller [3], "interpretability is the degree to which a human can understand the cause of a decision." To make sure that the human-understood causes are aligned with model's internals, we also consider faithfulness. As defined in Jacovi and Goldberg [4], "a faithful interpretation is one that accurately represents the reasoning process behind the model's prediction." Moreover, we define a reliable NLP system to be a system that produces consistent and accurate responses. Therefore, reliability is closely related to metrics such as high accuracy, low hallucination percentages, and consistency.

The importance of developing such systems cannot be overstated, as they are crucial for ensuring the responsible deployment of AI technologies across diverse applications, from healthcare and finance to autonomous systems and beyond. To illustrate the importance of trustworthy AI, consider the following scenario. Imagine a hospital that implements an advanced LLM assistant to help physicians diagnose patients according to their symptoms, medical history, and test results. This AI assistant processes vast amounts of medical literature, clinical data, and patient records to suggest potential diagnoses and treatment plans, reaching high accuracy in academic experiments. However, when implemented in real life, the AI assistant recommends invasive surgeries to a specific patient with mild symptoms. In this case, without understanding the reasons for the AI's decision, the doctors will find it hard to trust its recommendation as misguided choices would result in serious consequences. Even though the system may reach higher accuracy compared to humans, it lacks controllability, raising concerns about its reliability in such high-stake scenarios. Therefore, developing reliable and trustworthy AI systems is of critical importance in real-life AI deployment.

Current approaches to enhance the trustworthiness and reliability of NLP systems focus on improving interpretability and increasing controllability. These methods can be broadly categorized into three main areas: intrinsically interpretable models, post-hoc visualizations and explanations, and language-based approaches. Intrinsically interpretable models aim to design AI systems that are inherently transparent in their decision-making processes such as linear regression models, often sacrificing some performance for clarity. Post-hoc visualization techniques and explanations attempt to elucidate the internal workings of complex models after they have been trained, offering insights into feature importance and decision boundaries. Language-based approaches, such as Chain-of-Thought (CoT) prompting, leverage the linguistic capabilities of LLMs to generate step-by-step reasoning, providing a more human-interpretable explanation of the model's thought process.

Despite these advancements, significant challenges persist in establishing trustworthy and reliable NLP systems, particularly in the context of LLMs. Firstly, it is difficult to explain LLM decisions using traditional approaches due to the large number of parameters and complex structure. The vast number of parameters often leads to spurious correlations, obscuring the true causal relationships that are crucial for meaningful explainability. In other words, it is difficult to find the

truly causal factors that caused a decision as many parameters could be correlated. Moreover, the high dimensionality of LLMs renders traditional visualization approaches impractical, as the number of neurons and connections far exceeds what can be effectively represented using current techniques. Secondly, it is hard to control the LLM's behaviors because of the generative nature of such models, which inevitably leads to hallucinations (plausible but factually incorrect information) and results in user skepticism. This distrust further leads to unreliability, as LLM outputs can be highly context-dependent and sometimes unpredictable. The lack of intuitive methods for visualizing LLM decision processes compounds this issue, leaving users with limited means to scrutinize or validate the model's outputs.

In light of these challenges, this thesis aims to address the fundamental question: How can we improve current NLP systems, particularly LLMs, to become reliable and trustworthy without model revisions or performance harm? By exploring novel methodologies for model interpretation, developing robust techniques for mitigating hallucinations, and proposing innovative approaches to enhance user trust, this research seeks to bridge the gap between the remarkable capabilities of modern LLMs and the critical need for their responsible and transparent deployment.

The remaining part of the introduction is structured as follows: We first introduce the current challenges in the field that motivated our research in Section §1.1. Then, we list out the main contributions of the research papers in Section §1.2, where we clearly summarize the main objectives, methodologies, and findings for each research project. Finally, we illustrate the structure and organization of the thesis in Section §1.3.

1.1 Motivations

The pursuit of trustworthy and reliable artificial intelligence systems has become increasingly critical as Large Language Models (LLMs) continue to advance and permeate various aspects of society. However, the unique characteristics of LLMs present formidable challenges to achieving this goal, necessitating innovative approaches and methodologies.

- **Complex Structure of LLMs**: The sheer scale and complexity of LLMs pose significant obstacles to understanding their internal mechanisms. Modern LLMs often contain billions of parameters with complicated interconnections. Firstly, this complexity renders traditional visualization techniques inadequate, as they struggle to capture and represent the multifaceted decision-making processes occurring within these models. Secondly, as the factors are highly correlated, it is difficult to uncover the truly causal reasons for an LLM decision. As the goal of interpretability is to uncover the causal chains that lead to specific predictions or generations, the inability to distinguish between mere correlations and genuine causation hampers our capacity to build trust in these systems and the explanations derived by conventional methods.

- **Open-ended Generative Nature of LLMs**: The open-ended generative capabilities of LLMs, while immensely powerful, introduce unique challenges in terms of trustworthiness and reliability. The potential for hallucinations erodes user confidence and poses risks in critical applications. Moreover, the difficulty in controlling LLM outputs makes them less reliable in real-world scenarios where consistency and predictability are critical.

In light of these challenges, the overarching objective of this thesis is to address a fundamental question: **How can we enhance current NLP systems, particularly LLMs, to become more trustworthy and reliable without necessitating model revisions or compromising performance?** This ambitious goal seeks to develop methodologies and frameworks that can be applied to existing models, thereby offering practical solutions that can be implemented without the need for resource-intensive retraining or architectural changes. To pursue this objective, we articulate three key research questions that guide our investigation:

RQ1: How can we interpret the decisions of language models in a causal manner, identifying the true reasons behind their outputs rather than relying on correlational parameters? In other words, how can we disentangle the complex web of interactions within LLMs to reveal the causal pathways that lead to specific predictions.

RQ2: Can we interpret language model predictions using a user-friendly unit of explanation that is intuitive and easily comprehensible to end-users, instead of traditional heatmaps or word importance scores?

RQ3: How can we enhance LLM controllability through minimal interventions, improving the reliability and predictability of LLM outputs without requiring extensive modifications to the underlying model architecture or the training process?

By addressing these research questions, this thesis aims to make contributions to the field of trustworthy and reliable NLP systems. The outcomes of this research have the potential to not only advance our theoretical understanding of LLMs but also to provide practical tools and methodologies for improving their trustworthiness and reliability in real-world applications. Ultimately, this work seeks to pave the way for the responsible and effective deployment of LLMs, ensuring that these powerful technologies can be harnessed to their full potential while maintaining the trust and confidence of users and stakeholders.

1.2 Major Contributions

In the following sections, we present our contributions to the field of trustworthy and reliable NLP systems, particularly focusing on Large Language Models (LLMs). To increase the reliability of LLMs with minimal revisions to the LLM itself, we propose methodologies covering various domains, including using interpreter models to explain LLM's internals, verbalizing the LLM's decision processes using prompt techniques, and grounding LLM generations using retrieval augmented generation (RAG). Each contribution addresses specific aspects of our research questions, advancing our understanding and providing practical solutions to enhance the interpretability, explainability, and controllability of LLMs.

- To address RQ1 and RQ2, we introduce the High-Impact Concepts (HI-Concept) framework in Chapter 3, an innovative approach for deriving causal and user-friendly explanations for Large Language Models (LLMs). This method extracts predictive high-level features, *i.e.,* concepts, from the model's hidden layer activations and optimizes for concepts whose existence substantially influences output predictions. Our approach moves beyond correlational explanations, addressing the challenges posed by the compositional

nature of language. Extensive experiments on real and synthetic tasks demonstrate that High-Impact Concepts achieve superior results in predictive impact, explainability, and faithfulness while maintaining the original model performances compared to baseline methods. By using concepts as a user-friendly unit of explanation, this framework significantly advances our ability to interpret LLM decisions causally.

- To address RQ2 and RQ3, we present the Verify-and-Edit framework in Chapter 4, a novel approach to enhance the factual correctness and controllability of Large Language Models (LLMs) in knowledge-intensive tasks. This method extends the Chain-of-Thought (CoT) prompting technique by incorporating a post-editing mechanism that leverages external knowledge sources. By verifying and editing the generated reasoning chains, our framework significantly improves the factual accuracy of LLM outputs without requiring model modifications. Experimental results demonstrate notable improvements in accuracy across various open-domain question-answering tasks when applied to GPT-3. This contribution not only enhances the reliability of LLM-generated content but also provides a user-friendly approach to increase model controllability, thereby fostering greater trust in LLM applications.

- To further address Research Question 3, we introduce Chain of Knowledge (CoK) in Chapter 5, a framework designed to further enhance LLM controllability and reduce hallucinations in knowledge-intensive tasks. Compared to Verify-and-Edit that only uses textual information, CoK dynamically incorporates grounding information from heterogeneous sources, including both unstructured and structured knowledge bases. The framework operates in three stages: reasoning preparation, dynamic knowledge adapting, and answer consolidation. A key innovation is the adaptive query generator, which enables access to diverse knowledge sources through various query languages. By progressively correcting rationales and leveraging reliable factual information, CoK significantly improves the factuality and consistency of LLM outputs. Experimental results demonstrate consistent performance improvements across different domains, showcasing CoK's effectiveness in enhancing LLM controllability and reliability without necessitating model revisions.

- To further address RQ3 in a multimodal context, we present a comprehensive survey on Retrieving Multimodal Information for Augmented Generation in

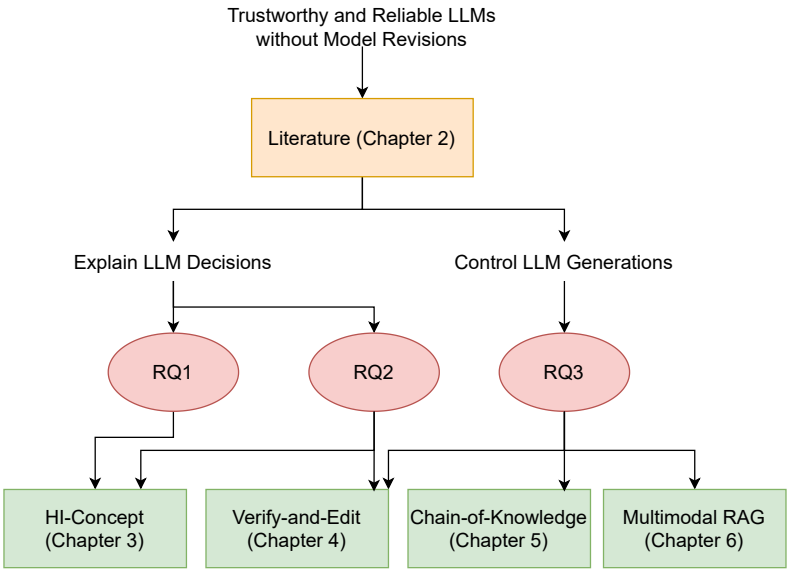


FIGURE 1.1: Overview of the Thesis.

Chapter 6. This survey synthesizes and analyzes the emerging trend of using multimodal information to enhance the controllability of Large Language Models (LLMs). By examining methods that incorporate diverse modalities such as images, code, tables, graphs, and audio, we provide a unified perspective on the stages and approaches for integrating multimodal knowledge. The survey highlights how these techniques can improve factuality, reasoning, interpretability, and robustness in LLM outputs. This contribution offers valuable insights into the current state of multimodal augmentation for LLMs, providing researchers and practitioners with a foundation for developing more controllable and reliable multimodal AI systems.

1.3 Outline of the Thesis

In this section, we provide a brief overview of the structure of this thesis, which is illustrated in Fig. 1.1.

- Chapter 2 presents a comprehensive literature review on NLP interpretability methods and approaches to increasing the trustworthiness of Large Language

Models (LLMs). This includes an examination of techniques for reducing hallucinations and enhancing LLM groundedness, providing a solid foundation for the research presented in subsequent chapters.

- To disentangle causal explanations and increase user trust for LLMs, Chapter 3 introduces the High-Impact Concepts framework. This novel approach extracts causal and user-friendly explanations for LLM decisions by identifying influential features from hidden layer activations. The chapter details the methodology and presents experimental results demonstrating improvements in predictive impact, explainability, and faithfulness.

- To increase the reliability of LLM systems and reduce hallucinations, Chapter 4 describes the Verify-and-Edit framework. This method enhances the factual correctness of LLM outputs in knowledge-intensive tasks by post-editing Chain-of-Thought reasoning chains using external knowledge sources. The chapter presents the framework's implementation and its performance improvements in open-domain question-answering tasks.

- To further increase LLM reliability by grounding the generations in multiple knowledge formats, Chapter 5 briefly introduces Chain-of-Knowledge (CoK) as a follow-up extension. This framework dynamically incorporates grounding information from heterogeneous sources to improve LLM controllability and reduce hallucinations. The chapter outlines the three-stage process and highlights the adaptive query generator for accessing diverse knowledge sources.

- To improve LLM reliability by grounding their generations in multimodal contexts, Chapter 6 presents a comprehensive survey on Retrieving Multimodal Information for Augmented Generation. This chapter reviews methods for incorporating various modalities to enhance LLM capabilities, providing insights into improving factuality, reasoning, interpretability, and robustness in multimodal AI systems.

- Chapter 7 concludes the thesis, summarizing the key findings and contributions. It also discusses future research directions and potential applications of the developed frameworks in advancing trustworthy and reliable NLP systems.

Chapter 2

Literature Review

To better understand the current landscape of trustworthiness research, we now present literature reviews that survey approaches that advance the trustworthiness and reliability of LLMs (or AI in general). We divide the approaches based on their motivations into 2 sections: explaining LLM's decisions, making LLM generations more factual, and making LLM predictions more controllable.

2.1 Explaining LLM's decisions

Explainable AI (XAI) is an established field of study that attempts to explain the Artificial Intelligence systems' internal decision processes. To define it more formally, Miller [5] gives the following definition: "Interpretability is the degree to which a human can understand the cause of a decision." Kim et al. [6] defines it as: "Interpretability is the degree to which a human can consistently predict the model's result." From the definitions, we can see that XAI is a critical aspect of the deployment of models in real-life scenarios.

Illustrated in Doshi-Velez and Kim [7], **causality** is an important consideration for interpretable ML. It implies whether the predicted change in output due to a perturbation will occur in the real system. However, it is an often overlooked attribute in XAI methods. Motivated by RQ1, we survey XAI methods in two families below: general explanability (non-causal) methods and causal explanability methods.

2.1.1 General Explainability Methods

Below, we will discuss XAI methods either for AI in general or with a specialization in NLP.

2.1.1.1 XAI methods for AI in General:

The need for explainable AI has grown with the increasing complexity and widespread adoption of AI systems across various domains. Doshi-Velez and Kim [7] provide a comprehensive overview of the motivations behind XAI, emphasizing its importance for building trust, enabling system debugging, and ensuring compliance with legal and ethical standards. Overall, explainability methods can be classified based on different criteria. We list the commonly used criteria below:

- **When the explainability method is applied:**
  - *Intrinsic:* Intrinsic approaches design simple model structures that can be directly interpreted. Rudin [8] argues for the superiority of such intrinsically interpretable models over complex "black-box" models used with post-hoc explanations. Examples of such interpretable models include decision trees [9], rule-based systems [10], and sparse linear models. These models offer transparency in their decision-making process, allowing for a direct interpretation of their predictions.
  - *Post-hoc:* Post-hoc methods are those applied to explain a model's decision after the model has been trained. They are particularly used for more complex models, such as deep neural networks. Ribeiro et al. [11] introduced LIME (Local Interpretable Model-agnostic Explanations), which approximates the behavior of complex models locally around specific predictions using simpler and more interpretable models. This approach has gained significant attention because of its model-agnostic nature and the ability to provide intuitive explanations. Another influential method in the field of XAI is SHAP (SHapley Additive exPlanations), proposed by Lundberg and Lee [12]. SHAP unifies several previous approaches to feature importance, providing a framework based on game theory to assign importance values to each feature for a particular prediction.

- **What format the derived explanation is:**
  - *Feature visualizations:* Simonyan et al. [13] introduced saliency maps for explaining image classification models, highlighting the regions of an input image that are most critical for the model's decision. This approach has then been extended and refined by numerous researchers, leading to more sophisticated visualizations for diverse types of neural networks.
  - *Counterfactual explanations:* Local data points consist of another important format of XAI explanations. Wachter et al. [14] proposed a framework for generating counterfactual explanations, which provide insight into how a model's prediction would change if certain input features were different. This approach offers actionable insights in domains such as finance and healthcare.
  - Besides the formats introduced above, we can also derive explanations in the form of *feature summary statstic* such as feature importance scores, *model internals* such as specific neurons pinpointed to be responsible for a decision, and *interpretable models* that approximate the AI system's behaviors.
- **What models the explanations are for:**
  - *Model-agnostic* explanations can be used for any ML model in a post-hoc way. Ribeiro et al. [15] analyzes the motivations and strengths of such methods. Their key advantage is the provided flexibility in the choice of models and representations. The previously introduced LIME [11] is an example of such method.
  - *Model-specific* tools are for a specific model class. For example, explanations using the sparse linear models are always model-specific.
- **What the explanations explain:**
  - *Global* explanations explain how the model functions as a whole. For example, it explains the "mindset" of the model and how it thinks. For example, concept-based methods such as CAV [16] (Concept Activation Vectors) discover global "concepts" that are key elements of how the model functions. It discovers concept vectors that divide the decision boundaries of a model's classifications.

– *Local* or *instance-specific* explanations explain why the model made a specific decision on an individual instance. For example, counterfactual explanations [14] derive perturbations to local instances that would lead to different decisions.

Specifically, we dive into a group of methods that are most closely related to the research presented in Chapter 3, concept-based methods, which are a family of post-hoc and model-specific methods that discover global explanations. Concept-based methods have been a popular group of interpretability methods as they derive user-friendly, high-level concepts as explanations. Kim et al. [16] first introduces Concept Activation Vectors (CAV) that align with user-defined concepts. It trains a linear classifier between examples and random counterexamples and takes the orthogonal vector to the decision boundary as concepts. Koh et al. [17] further learns high-level concepts and experiments with how the user could interact to edit concepts during test time. The most recent method, ConceptShap [18], discovers concepts in the intermediate layer with a bottleneck-shaped extraction model and proposes an adapted Shapley value metric to evaluate completeness scores.

Current Challenges: Traditional explainable AI (XAI) methods face significant challenges in uncovering true causal relationships, as they primarily identify correlations rather than causation. While these techniques excel at interpreting model behavior by highlighting feature importance or decomposing predictions, they often fail to distinguish between spurious associations and genuine causal mechanisms. This limitation stems from their reliance on statistical patterns within training data, which may not reflect underlying causal structures. The research gap lies in bridging the divide between post-hoc interpretability and causal reasoning. Consequently, practitioners risk drawing misleading conclusions, especially in high-stakes domains where understanding causality is critical.

For example, as we will show in Chapter 3, because existing concept-based methods do not differentiate between correlational and causal information, their performance on NLP tasks is problematic: the concepts discovered often have little impact on the predictions of the final model. Especially in complex transformer models with stronger confounding effects brought by pretraining, their performances may further decrease.

2.1.1.2 XAI methods specialized in NLP:

In the NLP domain, the need for explainable AI is particularly acute due to the complexity and ambiguity inherent in human language. The rise of LLMs further intensified this need, as these models are black-box in nature and often make decisions based on intricate patterns derived from vast amounts of text data.

Attention mechanisms, introduced by Chorowski et al. [19] for speech recognition and later adapted for transformers, have become a cornerstone of explainability in NLP. Although originally designed to improve model performance, attention weights can also pinpoint which parts of the input text the model focuses on when making predictions. However, [20] caution against overinterpreting attention weights as explanations, demonstrating that attention weights may not always correlate with gradient-based measures of feature importance. Addressing the challenge of faithfulness in explanations, Wiegreffe and Pinter [21] further introduced attention regularization, which aims at aligning attention weights more closely with model behavior. This approach seeks to make attention-based explanations more reliable and indicative of the model's true decision-making process.

For text classification tasks, Wallace et al. [22] introduces input triggers, which are a short sequence of words that can successfully trigger the target. It discovers such triggers by Input Reduction, a method that iteratively removes the least important words from the input while maintaining the model's prediction. This approach helps identify the minimal set of words crucial for the model's decision. Another notable line of work derives feature importance scores, which highlight the degree of importance of each word or feature in an LLM decision. Croce et al. [23] proposes using Layerwise Relevance Propagation on a Kernel-based Deep Architecture to generate transparent, analogy-based explanations for NLP system decisions, demonstrating effectiveness in tasks like question classification and semantic role labeling while addressing the need for accountability in increasingly pervasive NLP systems.

In the context of LLMs, the possibility of the LLM articulating the reasons before generating the final output has also provided an alternative explanation unit, which is a textual rationale. The trend started with the Chain-of-Thought or CoT [1] method, which is a prompting method for improving the reasoning abilities

of LLMs, which enables LLMs to decompose complex problems into multiple intermediate steps. CoT provides interpretability and has been proven to be more capable of solving complex problems than standard prompting methods. Besides the improved performance, the generated textual rationales also make the LLMs more transparent and interpretable.

For more structured explanations in NLP, Camburu et al. [24] proposed e-SNLI, a dataset of human-written explanations for natural language inference. This work paved the way for training models to generate human-like explanations alongside their predictions, enhancing interpretability. Rajani et al. [25] introduces the Common Sense Explanations (CoS-E) dataset and the Commonsense Auto-Generated Explanation (CAGE) framework, which uses human-provided and automatically generated explanations to improve the performance of deep learning models on commonsense reasoning tasks, demonstrating significant improvements in natural language understanding and generalization. To train models using these collected rationales, Ling et al. [26] proposes using answer rationales as indirect supervision for program learning to solve algebraic word problems, demonstrating improved explainability and effectiveness in natural language processing for arithmetic tasks. Some researchers design inherently explainability-aware architectures, which try to alleviate the black-box nature of language models.

Current Challenges: In conclusion, while significant progress has been made in developing XAI techniques for both general AI and NLP-specific applications, many challenges remain. These include improving the faithfulness and robustness of explanations, developing methods that scale to very large models, and bridging the gap between technical explanations and human understanding. As AI systems become increasingly integrated into decision-making processes in various domains, the continued advancement of XAI research will be crucial to ensure their responsible and reliable deployment.

2.1.2 Causal explainability methods

Causal explainable AI (XAI) methods have emerged as a powerful approach to understanding and interpreting complex AI systems. By focusing on causal relationships rather than mere correlations, these methods aim to provide more robust

and meaningful explanations.

2.1.2.1 Causal XAI methods for AI in General

The integration of causal reasoning into explainable AI has gained significant attention in recent years, driven by the need for more reliable and actionable explanations. Pearl [27] laid the foundation for this line of research, emphasizing the importance of causal models in AI and introducing the "ladder of causation" framework, which distinguishes between associational, interventional, and counterfactual reasoning.

Building on top of this foundation, Chattopadhyay et al. [28] proposed a framework for developing causal neural network attributions. Based on structural causal models (SCM), they aim to identify the causal effect of input features on model predictions, providing a more nuanced understanding of feature importance than traditional correlation-based methods.

For tabular data, Karimi et al. [29] surveys algorithmic recourse methods through contrastive explanations. Such methods provide actionable recommendations for changing an unfavorable outcome, grounded in causal reasoning about the effects of interventions on model predictions.

In the domain of reinforcement learning, Madumal et al. [30] proposed a framework to derive causal explanations of the behavior of model-free reinforcement learning agents. By learning structural causal models (SCM) during reinforcement learning, their approach generates explanations based on counterfactual analysis of the causal models.

For the domain of computer vision, Goyal et al. [31] introduced counterfactual visual explanations. They hope to find minimal changes to an input image that would alter the model's prediction, which then offers intuitive and causally-grounded explanations for image classification decisions.

To develop post-hoc explanations for neural networks, Harradon et al. [32] attempt to intervene in an unsupervised way on the hidden space by constructing several even-spaced Variational Autoencoders (VAEs) throughout a CNN, but they only

train with a reconstruction loss instead of explicitly optimizing for impact. In addition, there have been attempts to generate counterfactual inputs using disentangled VAEs, such as [33].

Another line of work uses probing and Causal Mediation Analysis (CMA) to explain black-box models. *Probing* methods [34–36] train an external model - a *probe* - to predict properties from the latent representations. However, it suffers from inherent flaws [37, 38], such as poor generalization. However, subsequent work [37, 38] shows that such methods generalize poorly to unseen samples and features may not be utilized in the original model's prediction task. To further investigate the causal effects of the features learned from probing, Elazar et al. [36] assess the influence of a causal intervention by removing a feature. However, subsequent work [37] shows that such methods generalize poorly to unseen samples. Moreover, as Belinkov [38] points out, the disconnect between the probing model and the original model may result in the properties not being utilized in the original model's prediction task. *Causal Mediation Analysis (CMA)* [39, 40] measures the change in output following a counterfactual intervention in an intermediate variable, or mediator. Both methods can be viewed as supervised concept discovery algorithms.

Current challenges: Despite significant advances in causal XAI methods, several challenges and research gaps remain. The application of causal reasoning in explainable AI often struggles with scalability issues when dealing with high-dimensional data and complex model architectures. Many current approaches rely on simplified causal models that may not fully capture the intricate dependencies in modern AI systems. There is also a notable gap between theoretical causal frameworks and their practical implementation, particularly in domains with limited domain knowledge for constructing accurate causal graphs. The evaluation of causal explanations presents another challenge, as there are few standardized metrics to assess their quality and utility. Furthermore, most existing methods focus on static causal relationships, overlooking the dynamic nature of many real-world systems where causal relationships evolve over time. Research is also needed to address the trade-off between the sophistication of causal explanations and their interpretability for non-expert users. Methods such as mediation analysis heavily rely on human-constructed features, which requires expertise and creates limitations. Thus, it may be beneficial to develop unsupervised explanation features. Finally, there remains limited work on integrating causal XAI approaches with

other explanation methods to provide more comprehensive and robust explanations across different AI contexts and applications.

2.1.2.2 Causal XAI methods specialized in NLP

In the field of NLP, causal XAI methods have been adapted to address the unique challenges posed by textual data and language models. According to Feder et al. [41], causality shows a promising path forward for NLP research, which can offer insights into the model's inner workings.

During pretraining, Feder et al. [42] introduced CausaLM, a framework for injecting causal knowledge into language models during pre-training. This approach aims to improve both the performance and interpretability of language models in causal inference tasks, demonstrating the potential for integrating causal reasoning directly into model architectures.

Addressing the challenge of identifying causal relationships in text, Li et al. [43] proposes a conditional text generation framework that posits sentential expressions of possible causes and effects. They first developed CausalBank, a large-scale collection of English sentences expressing causal patterns. Then, they use CausalBank to perform continued training to support causal reasoning.

In the context of bias mitigation, Vig et al. [40] developed a causal mediation analysis (CMA) framework to examine gender bias by changing pronouns in the input. By tracing the flow of information through the model, their method identifies specific components responsible for generating biased outputs, offering insights for targeted debiasing interventions. Prabhakaran et al. [44] develops methods to uncover and quantify various types of biases in word embeddings and downstream NLP tasks. This work highlights the importance of XAI techniques in ensuring fairness and ethical deployment of NLP systems.

Most causal XAI methods in NLP seek to develop counterfactuals as a form of explanation, Alvarez-Melis and Jaakkola [45] use a Variational Autoencoder (VAE) to generate counterfactuals and conduct causal analysis. Veitch et al. [46] conduct stress tests by perturbing input words. Wu et al. [47] construct a low-cost counterfactual generator for downstream applications. Such counterfactual explanations,

however, require extra caution to hold rigorously in causality, as causal and correlational relationships exist among input features and we cannot explicitly obtain such casual graphs or correlations in practice. To overcome this challenge, we propose to perturb on the intermediate hidden layer, thus assuming the independence between latent concepts.

Current challenges: While significant progress has been made in developing XAI techniques for both general AI and NLP-specific applications, many challenges remain. These include improving the faithfulness and robustness of explanations, developing methods that scale to very large models, and bridging the gap between technical explanations and human understanding. As AI systems become increasingly integrated into decision-making processes in various domains, the continued advancement of XAI research will be crucial to ensure their responsible and trustworthy deployment.

2.2 Making LLM Generations More Factual

The rise of Large Language Models (LLMs) has revolutionized natural language processing, enabling impressive performance across a wide range of tasks. However, a significant challenge that has emerged is ensuring the factual accuracy of LLM-generated content. Because of their vast vocabulary and open-generative nature, hallucination has become a prominent concern for LLMs, which has drawn significant attention from research communities. The decoding process of LLMs is auto-regressive, which unavoidably makes it output nonfactual content without controlled generation [48, 49]. This issue is commonly referred to as **hallucinations** [50].

Reducing Hallucinations in Text Generations: One of the primary approaches to enhancing factuality in LLMs is by using *external knowledge sources*. Lewis et al. [51] introduced the Retrieval Augmented Generation (RAG) model, which combines a neural retriever with a sequence-to-sequence model. RAG retrieves relevant documents from a large corpus and conditions its generations on these documents, leading to more factual and specific outputs. Building on this work, Borgeaud et al. [52] developed RETRO (Retrieval-Enhanced Transformer),

which uses chunked cross-attention to incorporate information from retrieved neighbors efficiently, demonstrating improved performance on language modeling benchmarks while maintaining factual consistency.

Another line of research focuses on improving the **internal knowledge representation** of LLMs. Guu et al. [53] proposed REALM (retrieval augmented language model), which is a method for pre-training language models with a latent knowledge retriever. This approach allows the model to learn to retrieve and use relevant information during pre-training, leading to improved performance on knowledge-intensive tasks. Similarly, Roberts et al. [54] combined T5 with SSM (Salient Span Masking), which is a pre-training approach that encourages the model to focus on and retain factual information by masking salient spans of text. Results show that this method scales well with model size and performs competitively with models that have access to external knowledge.

Fact-checking and verification strategies have been explored as another approach to enhancing LLM factuality. Thorne et al. [55] introduced the FEVER dataset and task, which involves verifying claims using evidence from Wikipedia. This work has spurred further research into automated fact verification systems that can be integrated with LLMs. Building on this, Nakano et al. [56] developed WebGPT, a model trained to use web browsing to answer questions. WebGPT demonstrates improved factual accuracy by leveraging real-time web information and providing citations for its responses. Recent work has also focused on developing more robust evaluation metrics for factual consistency. TRUE [57] implements a group of factual consistency metrics. The metrics are for existing tasks where the authors have manually collected factual consistency annotations.

Addressing the challenge of factual consistency in dialogue systems, Shuster et al. [58] explores the use of neural-retrieval-in-the-loop architectures for knowledge-based dialogue, which has been shown to significantly reduce hallucinations in conversations. Similarly, Rashkin et al. [59] explored enhancing truthfulness in dialogues by introducing novel evaluation metrics to categorize responses based on their information content. They incorporated these metrics as supplementary data during the model's training phase. Subsequently, these added inputs served as style controls, guiding the model towards producing more accurate responses when deployed. This controlled output approach resulted in dialogue generation that demonstrated improved objectivity and reliability compared to standard methods.

Reducing Hallucinations in CoTs: In addition to the methods above for general hallucination alleviation during text generation. In reasoning-intensive tasks. Researchers has noted that the lack of supporting facts during the generation process of CoT could largely undermine the validity of the final answer [60]. Ye and Durrett [48] demonstrate that the accuracy of the final answers largely correlates with the factuality and consistency of the reasoning explanations. The commonly proposed methods to improve the factuality of the CoT reasoning process can be grouped into two categories: prompt engineering and result calibration.

Prompt engineering methods are usually applied to guide LLMs to generate better intermediate reasoning explanations. *ReAct* [61], which is the most comparable to our work, synergizes reasoning and acting in LLMs, where reasoning steps help the model induce and update actions, while action steps allow the model to consult additional information from Wikipedia for a factuality check. Compared to *ReAct*, we generate more natural and conversational CoTs for better interpretability and easier learning. As such, our framework requires a much shorter prompt to learn. Press et al. [62] propose *self-ask* by instructing the LLM to explicitly ask itself (and then answer) follow-up questions before answering the initial question. One natural way of solving a complex problem is to decompose the problem into subproblems and solve them sequentially. Zhou et al. [63] adopt the idea and propose *least-to-most* prompting. However, both *self-ask* and *least-to-most* prompting still rely on repetitively retrieving internal knowledge learned by the LLM instead of connecting to external knowledge. Thus, their ability to improve factuality is limited.

Result calibration functions on the output of the LLMs. Ye and Durrett [48] train a calibrator to calibrate the weights of the final answers based on the factuality and consistency of the generated explanations, which efficiently improves the results. The decoding method in CoT is naive greedy, which simply outputs the next token with the highest probability. Wang et al. [2] propose a *self-consistency* decoding method. This technique generates multiple potential reasoning paths and then evaluates the consistency of these various pathways to determine the most reliable one, effectively filtering out inconsistent or less probable outcomes. *Selection-Inference (SI)* [64] framework is another state-of-the-art method that exploits LLMs as general processing modules. Out of all the methods, it is also the first to systematically improve the factual correctness of CoTs in order to predict

more accurately. It alternates between selection and inference to generate a series of interpretable causal reasoning steps leading to the final answer, which is proven to be efficient. However, it is not designed for answering open-domain or common-sense questions.

Existing Challenges: Despite these advancements, there are still challenges in making LLM predictions consistently factual across diverse domains and tasks. Maynez et al. [65] highlighted the persistent issue of hallucination in abstractive summarization, emphasizing the need for continued research in this area. Additionally, LLMs could struggle with up-to-date information and evolving facts due to temporal knowledge cutoffs.

In conclusion, while significant progress has been made in improving the factuality of LLM predictions through various techniques such as retrieval augmentation, improved pre-training strategies, and fact-checking integration, there is still substantial room for improvement. As LLMs continue to be deployed in increasingly critical applications, ensuring their factual accuracy remains a paramount concern for the NLP research community.

2.3 Making LLM Generations More Controllable

In recent years, the rapid advancement of Large Language Models (LLMs) has revolutionized natural language processing tasks, demonstrating remarkable capabilities in text generation, question answering, and complex reasoning. However, as these models grow in size and complexity, controlling their outputs has become an increasingly critical challenge. Beyond the factuality element, controllability stresses conformity to the user's intended goals, topics, styles, and contents.

Controlling LM Outputs: The concept of controlled text generation has been further refined through the development of specialized frameworks. Before the age of LLMs, Keskar et al. [66] introduced CTRL, a conditional transformer language model that allows for control over style, content, and task-specific behavior through the use of control codes. This approach enables a more precise steering of the generation process, allowing users to specify attributes such as domain, style, or

topics. To ensure consistent persona or style in language generation, Zhong et al. [67] proposed a new task towards persona-based empathetic conversations. They first collect a dataset for persona-based empathetic conversations and then propose CoBERT, a response selection dataset trained on this new task. It then allows for the generation of responses that are consistent with a given persona. Following this line of work, Roller et al. [68] developed BlenderBot, which incorporates persona consistency mechanisms, allowing for more controlled and coherent multi-turn dialogues.

In the realm of story generation and long-form text production, Fan et al. [69] introduced a hierarchical story generation model that allows for control over plot elements. Goldfarb-Tarrant et al. [70] further expanded on this with a plan-and-write approach, providing more fine-grained control over narrative structure and content.

Controlling LLM Outputs: To the end of controlling LLM outputs, one of the fundamental approaches to controlling LLM outputs is through prompt engineering. GPT-3 [71] demonstrated the power of in-context learning, where the model's behavior can be significantly influenced by the content and structure of the input prompt. Building on this, Reynolds and McDonell [72] conducted a comprehensive study on prompt design strategies, highlighting the importance of clear instructions, examples, and context in guiding model outputs.

Besides non-training approaches, a significant advancement in controllability came with the introduction of instruction tuning. Wei et al. [73] demonstrated that fine-tuning language models on a diverse set of tasks described via instructions leads to models that are better at following open-ended instructions. This approach, exemplified by models like InstructGPT [74] and FLAN-T5 [75], has significantly improved the ability of LLMs to adhere to specific user instructions, enhancing their controllability across a wide range of tasks.

The development of parameter-efficient fine-tuning methods has opened new avenues for customizing LLM behavior. Houlsby et al. [76] introduced adapter modules, which permit task-specific modifications without altering the entire pre-trained model. Hu et al. [77] further refined this with LoRA (Low-Rank Adaptation), offering an efficient method for adapting LLMs to specific domains or tasks, thereby enhancing controllability.

To eliminate the need of expensive re-training or fine-tuning, Plug-and-play language models, introduced by Dathathri et al. [78], offer a flexible approach to controlled text generation. This method allows for the adjustment of text generation towards specific attributes without retraining the entire model, providing a powerful tool for customizing LLM outputs.

In addition to controlling LLM outputs in style, recent work has also focused on controlling the ethical aspects of LLM generations. Solaiman and Dennison [79] introduced Process for Adapting Language Models to Society (PALMS), a methodological framework for modifying the conduct of language models to align with societal expectations. This iterative approach involves the creation and subsequent utilization of a specially curated dataset that embodies a predefined set of desired attributes. By fine-tuning models on this carefully constructed corpus, researchers hope to enhance the controllability of large language models, particularly in their adherence to ethical standards and social conventions. Bai et al. [80] presents an innovative approach to developing an LLM with enhanced ethical behavior through Reinforcement Learning from AI Feedback (RLAIF). This method initially generates outputs from a base model, followed by automated self-evaluation and refinement processes. The original model is then fine-tuned using these improved responses. Subsequently, the research employs a comparative analysis phase, wherein an AI evaluator assesses the relative quality of outputs from the refined model. This assessment data is utilized to train a preference model, effectively capturing AI-derived quality metrics. This novel methodology offers the potential for more precise control over AI behavior while significantly reducing the need for human-annotated data.

Existing challenges: While significant progress has been made in enhancing the controllability of LLM generations through techniques such as prompt engineering, instruction tuning, persona-based models, and plug-and-play approaches, there is still considerable room for improvement. As LLMs continue to be integrated into various applications, the ability to reliably and precisely control their outputs remains a critical area of research in the NLP community.

Chapter 3

Explaining Language Model Predictions with High-Impact Concepts

To encourage fairness and transparency, there exists an urgent demand for deriving reliable explanations for large language models (LLMs). One promising solution is concept-based explanations, *i.e.,* human-understandable concepts from internal representations. However, due to the compositional nature of languages, current methods mostly discover *correlational* explanations instead of *causal* features. Therefore, we propose a novel framework to provide impact-aware explanations for users to understand the LLM's behavior, which are robust to feature changes and influential to the model's predictions. Specifically, we extract predictive high-level features (concepts) from the model's hidden layer activations. Then, we innovatively optimize for features whose existence causes the output predictions to change substantially. Extensive experiments on real and synthetic tasks demonstrate that our method achieves superior results on predictive impact, explainability, and faithfulness compared to the baselines, especially for LLMs.

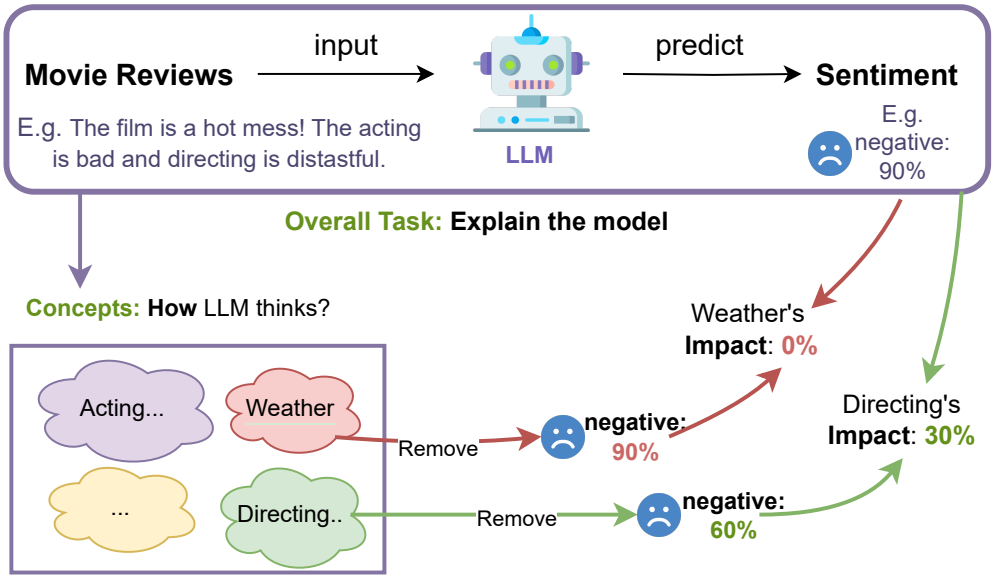


FIGURE 3.1: Illustration of concept-based explanations that result in high impact (green line) or not (red line) when explaining the LLMs in a sentiment classification task.

3.1 Chapter Background

Over the past few years, large language models (LLMs) have achieved tremendous progress, leading them to be widely applied in sensitive applications such as personalized recommendation bots and recruitment. However, Explainable AI (XAI) has not witnessed the same progress, making it difficult to understand LLMs' opaque decision processes [81]. Therefore, many users are still reluctant to adopt LLMs in high-stake applications due to transparency and privacy concerns. In this work, we aim to increase user trust and encourage transparency by deriving explanations that allow humans to better predict the model outcomes.

To understand what happens inside an LLM, previous studies [82] show that dense vector representations in high layers of a language model tend to capture semantic meanings that are useful for solving the underlying task. However, such vector representations are not understandable to humans. To solve it, concept-based explanations attempt to map the hidden activation space to human-understandable features. For example, Koh et al. [17] provides the concept bottleneck model, which first predicts an intermediate set of human-specific concepts, and then uses them to predict the target. As illustrated by purple boxes in Fig. 3.1, for the movie review

classification task, concept-based explanations are semantically meaningful word clusters [82] corresponding to abstract features such as "acting" and "directing".

However, existing concept-based methods do not consider of the *explanation impact* on output predictions, leading to inferior explanations. By *impact*, we mean the causal effect of removing a feature on output predictions [83, 84]. As Moraffah et al. [85] points out, these non-impact-aware methods derive correlational explanations that cannot answer questions about decision-making under alternative situations and are thus unreliable. An example is illustrated in Fig. 3.1. Due to the conventional expression "hot mess", the word "hot" often co-occurs with "mess", which is usually used to classify negative sentiment. Traditional concept-based methods that do not consider impact may falsely use the correlational feature "weather" (*i.e.*,, "hot") to explain why the model classifies something as negative. However, excluding the "weather" concept does not cause the output prediction to change at all, resulting in zero impact (red line). Thus, low-impact explanations such as "weather" are less valid as users cannot utilize them to consistently predict the model's behaviors when a feature changes.

To tackle this bottleneck and incorporate impact into traditional concept-based models, in this work, we propose High-Impact Concepts (*HI-concept*), a complete concept explanation framework with causal impact optimization (§3.3.2). Specifically, we design a *causal* loss objective, stemming from the treatment effects in the causality literature [86]. Moreover, previous causality evaluations [83, 87] primarily focused on assessing the causal effect via *local* (*i.e.*,, instance-level) change and *removal* intervention (*i.e.*,, eliminating words/concepts from the source), leading to potentially biased evaluation results. To this end, we further propose a novel *global* (*i.e.*,, model-level) accuracy change metric and *insertion* operation to effectively diagnose the causality measurement (§3.3.4).

As a result, our method can consistently prioritize more influential features (green line in Fig. 3.1) while disregarding correlational ones.

Extensive experiments with multiple language models, both established and newly proposed evaluation metrics, and rigorous human studies fully validate the effectiveness of *HI-concept* in finding high-impact concepts compared to baselines, especially for LLMs. Our contributions are summarized as follows[1]:

[1]Our codebase is available at `https://github.com/RuochenZhao/HIConcept`.

- To alleviate the problem of correlational explanations, we propose *HI-concept*, a framework for deriving explanatory features with high impacts by innovatively optimizing a causal objective.
- Towards comprehensive evaluations, we propose a theoretically grounded metric, namely reconstruction accuracy change, and devise an insertion study, which serves as a complement to the traditional removal intervention.
- Extensive experiments show that *HI-concept* is impactful, explainable, and faithful, with especially outstanding improvements on LLMs (*e.g.*,, improving the causal effect on accuracy from **2.83**% to **27.79**% on Llama-7B).

3.2 Preliminaries

We first introduce what concept-based explanations are, what properties they should satisfy, and our key baseline, concept bottleneck models.

3.2.1 Concept-based Explanations

Concept-based explanations is a well-established method [16–18] that extracts human-understandable concepts from the model's hidden space. As stated in Kim et al. [16], the activation space of an ML model can be seen as a vector space E_m spanned by basis vectors e_m which correspond to input features. Humans work in a different vector space E_h spanned by implicit vectors e_h corresponding to an unknown set of human-understandable concepts. Then, concept-based explanations $g : \mathbb{E}_m \rightarrow E_h$ aim to translate from high-level representations into task-relevant and human-understandable concepts.

Ideally, concept-based explanations should satisfy the following properties [7]. *Faithfulness:* The explanations can be able to accurately mimic the original model's prediction process [11]. *Causality:* When the feature is perturbed in real life, the output predictions should change accordingly. This causal impact ensures that explanations are reliable under alternative situations. *Explainability:* The explanations should be understandable to humans and able to assist users in real-life tasks. These three properties will be the guiding principles for our model design and evaluation.

3.2.2 Concept Bottleneck Models

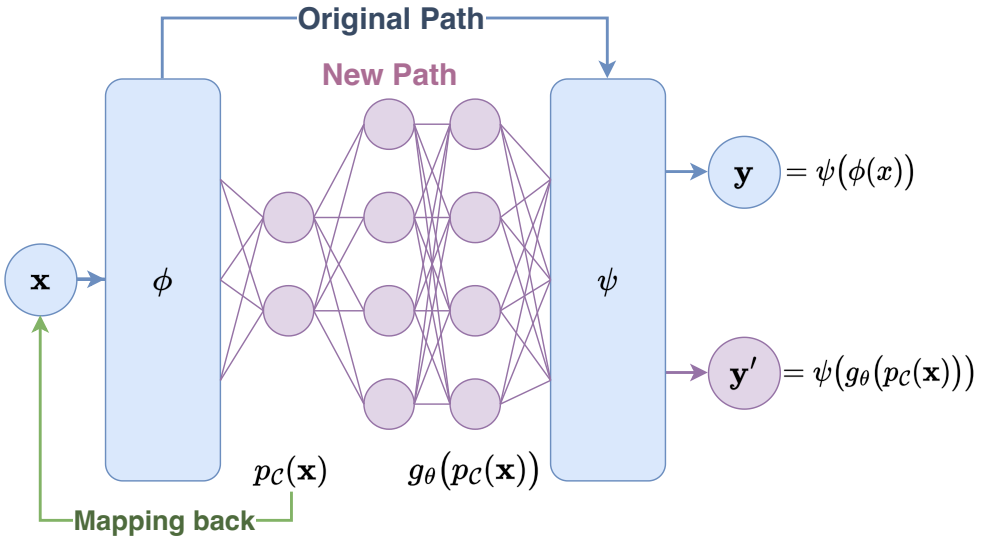


FIGURE 3.2: The overall concept generation process of a concept bottleneck model.

To derive concept-based explanations, one classic architecture is concept bottleneck models [18], shown in Fig. 3.2. The pretrained model f can be viewed as a composite of two functions, divided at an intermediate layer: $f = \psi \circ \phi$. After initializing the concepts $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \in \phi(\cdot)$ uniformly, $\phi(\mathbf{x})$ is encoded into concept probabilities $p_{\mathcal{C}}(\mathbf{x})$, calculated as $p_c^i(\mathbf{x}) = \mathrm{TH}((\phi(\mathbf{x})^\top \mathbf{c}_i), \beta)$[2] Then, the bottleneck-shaped network reconstructs $\phi(\mathbf{x})$ with a 2-layer perceptron g_θ such that $g_\theta(p_{\mathcal{C}}(\mathbf{x})) \approx \phi(\mathbf{x})$. Intuitively, the hidden space $\phi(\cdot)$ corresponds to the vector space E_m. The concept probability space $p_{\mathcal{C}}(\cdot)$ corresponds to the human-understandable space E_h. To train the concept model in an end-to-end way, two losses are used:

• *Reconstruction loss:* To faithfully recover the original model's predictions, a surrogate loss with cross-entropy (CE) is optimized[3]:

$$\begin{aligned} \mathcal{L}_{\text{rec}}(\theta, \mathcal{C}) &= \mathrm{CE}\Big(\psi\big(\phi(\mathbf{x})\big), \psi\big(g_\theta(p_{\mathcal{C}}(\mathbf{x}))\big)\Big) \\ &= -\sum_{b \in \mathcal{B}} \psi\big(\phi(\mathbf{x})\big)_b \log\big(\psi(g_\theta(p_{\mathcal{C}}(\mathbf{x})))_b\big). \end{aligned} \tag{3.1}$$

[2] TH is a threshold function that forces all inputs smaller than β to be 0.
[3] $\mathcal{B}$ is the set of class labels and $\psi(.)_b$ denotes the prediction score corresponding to label b.

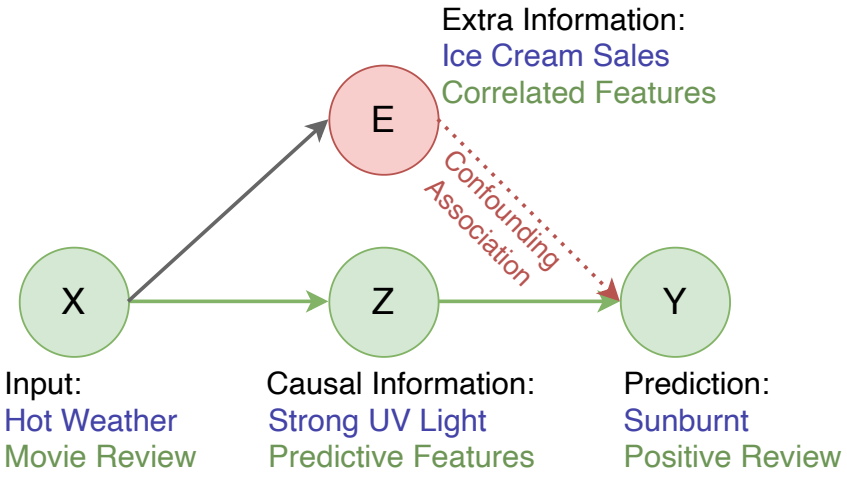


FIGURE 3.3: Illustration of the causal graph indicating the confounding association in explanation models. Blue is a real-life example. Green is the correspondence in a movie review classification task.

• *Regularization loss:* To make concepts more explainable, a regularization loss forces each concept vector to correspond to actual examples and concepts to be distinct from each other[4]:

$$\begin{aligned}\mathcal{L}_{\text{reg}}(\mathcal{C}) = &- \lambda_1 \frac{\sum_{i=1}^{n} \sum_{\mathbf{x}_t \in \mathcal{T}_{\mathbf{c}_i}} \mathbf{c}_i^\top \phi(\mathbf{x}_t)}{nN} \\ &+ \lambda_2 \frac{\sum_{i_1 \neq i_2} \mathbf{c}_{i_1}^\top \mathbf{c}_{i_2}}{n(n-1)}.\end{aligned} \tag{3.2}$$

3.3 Methodology

Then, we propose *HI-concept*, which aims to fill the current research gap on explanatory impact.

3.3.1 Defining Impact

As stated earlier, not considering impact could result in confounding and correlational explanations. The failure cases can be theoretically explained by causality analysis in Fig. 3.3. To achieve sentiment prediction Y , the hidden activation space

[4] $\mathcal{T}_{\mathbf{c}_i}$ as the set of top-k neighbors of $\mathbf{c}_i$

in pretrained LLMs consists of both correlated features E and predictive features Z. Although only Z truly affects the prediction Y, E and Z may be correlated due to the confounding effects brought by input X. However, a traditional concept mining model does not differentiate between E and Z and considers both as valid. Thus, it may easily use the confounding association as an explanation instead of the true causal path. The resulting concepts would be problematic as they do not facilitate a robust understanding of the model's behaviors.

To tackle this challenge, we enforce explanations to be predictive by considering their "impact". To formally define the *impact* of a feature, we utilize two important definitions in causal analysis: Individual Treatment Effect (ITE) and Average Treatment Effect (ATE), which measure the effect of interventions in randomized experiments [86]. Given a binary treatment variable T that indicates whether a *do-operation* is performed (*i.e.,*, perturb a feature), ATE and ITE are defined as the change in expected outcome with treatment $T = 1$:

$$\begin{aligned} \text{ITE}(x) := \; & \mathbb{E}[y|\mathbf{X} = x, \text{do}(T = 1)] \\ & - \mathbb{E}[y|\mathbf{X} = x, \text{do}(T = 0)]; \\ \text{ATE} := \; & \mathbb{E}[\text{ITE}(x)]. \end{aligned} \tag{3.3}$$

In our case, a concept $\mathbf{c}_i$ is discovered as a direction in the latent space, corresponding to a feature in the input distribution. As f is fixed, its prediction process is deemed deterministic and reproducible, allowing us to conduct experiments with treatments [17]. Therefore, we propose to remove a specific concept [83][5] as the do-operation and define *impact* I of a concept $\mathbf{c}_i$ on an instance $(\mathbf{x}, y)$ as:

$$\text{I}(\mathbf{c}_i, \mathbf{x}) = \mathbb{E}[y|\mathbf{X} = \mathbf{x}, \mathbf{c}_i = \mathbf{0}] - \mathbb{E}[y|\mathbf{X} = \mathbf{x}, \mathbf{c}_i = \mathbf{c}_i]. \tag{3.4}$$

3.3.2 Optimizing for Impact

In order to incorporate consideration for impact into the concept discovery process, we introduce two new losses to the original framework:

[5] We assume that, as the concept vectors coexist in the hidden embedding space, there is no causal relationship among the concepts $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ themselves.

• *Auto-encoding loss:* To guarantee that the intervened representations are still meaningful, we optimize an auto-encoding loss to learn a proxy task that reconstructs the hidden representations. With this loss, the concept model becomes Auto-encoder-like and can mimic a generation process of the real distribution of $\phi(\mathbf{x})$. Therefore, concept vectors can then be seen as key factors in the generation process of $\phi(\mathbf{x})$. Then, we can perform valid interventions on the concept vectors, such as the removal intervention. Formally:

$$\begin{aligned}\mathcal{L}_{\text{enc}}(\theta, \mathcal{C}) &= \text{MSE}\Big(\phi(\mathbf{x}), g_\theta(p_{\mathcal{C}}(\mathbf{x}))\Big) \\ &= \frac{1}{d}||\phi(\mathbf{x}) - g_\theta(p_{\mathcal{C}}(\mathbf{x})||_2^2.\end{aligned} \tag{3.5}$$

• *Causality loss:* Directly optimizing for causality is a challenging objective as causal impact is difficult to estimate during training. Therefore, we approximate impact (Eq. (3.4)) by randomly removing a set of concepts $\mathcal{S} \subseteq \mathcal{C}$ and calculating the expectation of impact on the training set. Then, we could disentangle concept directions that have a greater impact by optimizing the following loss:

$$\begin{aligned}\mathcal{L}_{\text{cau}}(\theta, \mathcal{C}) = -\sum_{\mathbf{c}_i \in \mathcal{S}} \sum_{\mathbf{x}_j \in \mathcal{D}} \Big| &\psi\Big(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)|\mathbf{c}_i = \mathbf{0})\Big) \\ &- \psi\Big(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)|\mathbf{c}_i = \mathbf{c}_i)\Big)\Big| \approx -|I_{\text{avg}}(\mathcal{C})|.\end{aligned} \tag{3.6}$$

As all inputs $\mathbf{x}_j \in \mathcal{D}$ are perturbed, the training dataset $\mathcal{D}$ serves both as the treatment group and the nontreatment group, ensuring no divergence.

Finally, the overall loss function becomes:

$$\begin{aligned}\mathcal{L}(\theta, \mathcal{C}) =& \mathcal{L}_{\text{rec}}(\theta, \mathcal{C}) + \mathcal{L}_{\text{reg}}(\mathcal{C}) \\ &+ \lambda_e \mathcal{L}_{\text{enc}}(\theta, \mathcal{C}) + \lambda_c \mathcal{L}_{\text{cau}}(\theta, \mathcal{C}),\end{aligned} \tag{3.7}$$

where λ_e, λ_c are the weights for the auto-encoding loss and the causal loss respectively. In practice, the hyperparameters require minimal tuning. Specifically, we recommend fixing $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ for regularizer loss in Eq. (3.2), and $\lambda_e = 1$ for reconstruction loss. The only hyperparameter to tune is λ_c, whose optimal level can be found within a few steps.

3.3.3 Visualizing Concepts via Impact

As a concept $c_i \in \phi(\cdot)$ is a hidden space vector, previous concept discovery methods face difficulties in mapping concept vectors to semantic meanings. They mainly relied on naively clustering the high-frequency words [18, 82]. To address this issue, we use established visualization techniques to translate it to human-understandable concepts (*i.e.*,, word clusters and highlights).

For models where the hidden representation is token-level, we simply use the individual token's concept probability $p_{\mathcal{C}}(x_i)$ as token importance scores. For models with sequence-level representations such as BERT, we employ the well-established transformer visualization method proposed in Chefer et al. [88] to map back from the [CLS] activation concepts to input tokens. As an adaption of Grad-CAM [89] to transformers, it visualizes classifications with layer-wise propagation, gradient backpropagation, and layer aggregation with rollout. As a result, for each sample $\mathbf{x}$ with tokens $x_1, \ldots, x_T$, we go from having only one concept similarity score $p_c^i(\mathbf{x})$ to a list of normalized token importance scores $s_1(\mathbf{c}_i), \ldots, s_T(\mathbf{c}_i)$. Therefore, we derive both global/model-level concepts (*i.e.*,, word clusters) and their corresponding local/instance-level explanations (*i.e.*,, token importance scores for an instance) that result in high impact. Both forms of generated explanations can complement each other while conforming to the model's 'mindset'.

3.3.4 Evaluating Impact of Concepts

Quantitatively, traditional causality evaluation metrics focus on local (*i.e.*,, instance-level) perturbations [87], which may be biased to global (*i.e.*,, model-level) performance evaluations. Thus, we innovatively propose *Recovering Accuracy Change (*Δ*Acc)*. Following the causality definition Doshi-Velez and Kim [7] and human intuition, if a concept $\mathbf{c}_i$ is a crucial factor used by the model to make predictions, omitting it should disrupt the ability to faithfully recover predictions. Formally, it is defined as:

$$\Delta\mathrm{Acc}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{c}_i \in \mathcal{C}} |\mathrm{Acc}(\mathcal{C}) - \mathrm{Acc}(\mathcal{C} \setminus \{\mathbf{c}_i\})|,$$

where Acc denotes the recovering accuracy [18].

Dataset	Train	Test	Label dim.	Avg. size
Toy (image)	48k	12k	15	(240, 240)
IMDB (text)	37.5k	2.5k	2	215
AG (text)	120k	7.6k	4	43

TABLE 3.1: A summary of the datasets.

Moreover, we follow previous work to use *Causal Concept Effect* (CACE) [83] to evaluate the causal effect of the set of concepts $\mathcal{C}$. Formally, it is defined as:

$$\begin{aligned}\text{CACE}(\mathbf{c}_i) := \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{test}}} |\psi\big(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j))\big) \\ - \psi\big(g_\theta(p_{\mathcal{C}\setminus\{i\}}(\mathbf{x}_j))\big)|; \\ \text{CACE}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{\mathbf{c}_i \in \mathcal{C}} \text{CACE}(\mathbf{c}_i)\end{aligned}$$

Qualitatively, existing evaluations mostly assess concepts' impact $\mathcal{C}$ via feature *removal* [83]. We argue that obtained concepts should also be generalizable to cases of *insertion*. Thus, we propose a novel insertion operation. Intuitively, when inserting explanation features one by one, gradual improvement of recovering accuracy should be observed, indicating incremental impact of each concept.

3.4 Experiment Setup

3.4.1 Datasets and Models

We test the effectiveness of our method with two standard text classification datasets: IMDB [90] and AG-news [91]. IMDB consists of movie reviews labeled with positive or negative sentiments, while AG-news is a dataset of news articles categorized into 4 topics. Table 3.1 gives a dataset summary.[6] We explain four classification models: (*i*) a 6-layer transformer encoder trained from scratch, (*ii*) a pre-trained BERT with finetuning, (*iii*) a pre-trained T5 model [92] with finetuning, (*iv*) 7B Llama [93] with in-context learning. The details on how we use these models are as follows:

[6]IMDB and AG-news are both licensed for non-commercial use.

Self-trained transformer: The self-trained transformer model used during text experiments follows a simple structure: the input text is truncated to max length 512 and passed to an embedding layer of dimension 200. Then, the embeddings are passed through a positional encoding layer with a dropout rate of 0.2. Then, 6 transformer layers follow with a hidden dimension of 200 and 2 heads. Finally, we mean-pool the transformed embeddings and pass through a linear classifier head. The linear outputs are activated with a Sigmoid function to produce class probabilities.

To train the transformer model, we use either the IMDB or AG-News dataset. We train for 10 epochs with a batch size of 128 and an Adam optimizer with learning rate $3e-4$. We also use a learning rate step scheduler with step size 1 and gamma of 0.95.

Dataset	AG-News	IMDB
LR	$5e-5$	$3e-4$
train BS	8	8
eval. BS	16	16
seed	42	42
optimizer	Adam	Adam
	betas $= (0.9, 0.999)$	betas $= (0.9, 0.999)$
	epsilon $= 1e-8$	epsilon $= 1e-8$
LR scheduler	linear	linear
warmup steps	7425	1546
training steps	74250	15468

TABLE 3.2: Hyperparameters for finetuning BERT model.

Pretrained and finetuned BERT model For AG-News, we take the fine-tuned version of bert-base-uncased model on huggingface: "fabriceyhc/bert-base-uncased-ag_news". For IMDB, we fine-tuned by ourselves on the bert-base-uncased model. The hyperparameters used for both finetuning are reported in §3.4.1, where LR stands for learning rate and BS stands for batch size.

The huggingface code and models are all licensed under Apache 2.0, which allows for redistribution and modification. Similarly, the codebase used to replicate the visualization method [88] and the baseline method [94] are licensed under the MIT license, which allows for redistribution of the code.

T5 and Llama As T5 and Llama are both generative models, when calculating impact, we simplify outputs by filtering to only the classification classes (e.g., words

"Positive", "Negative" for IMDB) and summing all other vocab probabilities as "Other".

For T5, we finetune on IMDB and AG-News separately using the same hyperparameters: max seq length of 512, learning rate of $3e-4$, weight decay of 0.0, adam epsilon of $1e-8$, warmup steps of 0, train batch size of 10, eval batch size of 10, num train epochs of 2, and gradient accumulation steps of 8.

The T5 model is licensed under Apache 2.0, which allows for redistribution and modification. For Llama, we use the 7B model licensed under GPL 3.0, which allows for redistribution and modification.

3.4.2 Metrics

We evaluate the explanation methods quantitatively and qualitatively with comprehensive metrics based on the three important considerations described in §3.2.1.
Faithfulness: To ensure that the surrogate model can accurately mimic the original model's prediction process, we evaluate whether the captured concept probabilities $p_{\mathcal{C}}(\mathbf{x})$ can recover the original model's predictions $\psi\big(\phi(\mathbf{x})\big)$ with the established metrics below:

(*i*) *Recovering Accuracy* (Acc): As defined in Yeh et al. [18], for the set of concepts $\mathcal{C}$, RAcc measures the prediction reconstruction accuracy using concept scores:

$$\text{RAcc}(\mathcal{C}) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{test}}} \mathbb{1}\Big(\psi\big(\phi(\mathbf{x}_j)\big) = \psi\big(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j))\big)\Big)$$

(*ii*) *Precision, Recall, F1*: To provide a thorough study, we also include common metrics including precision, recall, and F1 [95].

(*iii*) *Completeness*: As defined in Yeh et al. [18], completeness measures whether $\mathcal{C}$ is sufficient in recovering predictions. Denoting $\sup_g \mathbb{P}_{x,y\in\mathcal{D}_{\text{test}}}[y = \arg\max_{y'} \psi_{y'}(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)))]$ as the best accuracy by predicting the label just given the concept scores, and a_r as the accuracy of random prediction, completeness is formulated as:

$$\text{Completeness}(\mathcal{C}) = \frac{\sup_g \mathbb{P}_{x,y\in\mathcal{D}_{\text{test}}}[y = \arg\max_{y'} \psi_{y'}(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)))] - a_r}{\mathbb{P}_{x,y\in\mathcal{D}_{\text{test}}}[y = \arg\max_{y'} f_{y'}(x)] - a_r}$$

Causality is the key of the XAI model evaluation. As mentioned in §3.3.4, we use the CACE metric [83], a novel accuracy change metric (ΔAcc), and insertion operations to provide a more comprehensive overview. **Explainability.** With the concepts generating a high impact on predictions, we expect that it can allow end-users to better understand the model's decisions. We include visualizations and human studies to test it qualitatively.

Causality: To systematically evaluate causality, we conduct synthetic experiments, derive qualitative examples, draw trend graphs, and conduct human studies. In quantitative experiments, we use the following quantitative metrics:

(*i*) *Causal Concept Effect (CACE)*: As defined in Goyal et al. [83], CACE for a concept c is the change in prediction after removing it. Then, we compute the average CACE to evaluate a set of concepts $\mathcal{C}$:

$$\text{CACE}(\mathbf{c}_i) = \mathbb{E}\big[\psi\big(g_\theta(p_\mathcal{C}(\mathbf{x}_j))\big) - \psi\big(g_\theta(p_{\mathcal{C}\setminus\{i\}}(\mathbf{x}_j))\big)\big]$$

(*ii*) *Recovering Accuracy Change (ΔAcc)*: Doshi-Velez and Kim [7] state: "Causality implies that the predicted change in output due to a perturbation will occur in the real system". Therefore, if a concept $\mathbf{c}_i$ is a crucial factor used by the model to make predictions, omitting it should disrupt the ability to faithfully recover predictions:

$$\Delta\text{Acc}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{\mathbf{c}_i\in\mathcal{C}} |\text{RAcc}(\mathcal{C}) - \text{RAcc}(\mathcal{C}\setminus\{\mathbf{c}_i\})|$$

3.4.3 Baselines and Hyperparameters Used

For baselines, we use other unsupervised dimension reduction methods to discover concepts on the hidden space: (*i*) PCA [96] and K-means [97] are popular non-parametric clustering techniques that reduce high-dimensional datasets into key features to increase interpretability. (*ii*) β-TCVAE [94] is a disentanglement VAE method that explicitly considers causal impact while reducing dimensionality. (*iii*) ConceptSHAP [18] represents the traditional concept bottleneck models that do not consider impact.

For all concept experiments, the following parameters are universally applied as a selected default, which demonstrated better performances during experiments: For regularizer losses, $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$. In TH($\cdot, \beta$) function, threshold is set to be $\beta = 0.1 = \frac{1}{n}$, where n is the number of concepts selected. For the top-N neighborhood, $N = \frac{1}{4}$BS, where BS is the effective batch size, which we have set as 128 during the experiments. For the masking strategy, we always recommend masking random concepts with a probability of 0.2 as the optimal strategy, as masking the maximum concepts may lead to a highly uneven distribution of I($\mathcal{C}$) among discovered concepts.

As all dataset class sizes are small (2 in IMDB/toy or 4 in AG-News), the number of concepts is chosen to be 10 for all experiments. When the number of classes is larger, we recommend choosing a larger number of concepts to ensure a faithful reconstruction of the original input.

For training the concept model, we always use an Adam optimizer with a learning rate of $3e-4$. All models are all trained using 100 epochs. In the *HI-concept* models, causal loss is always turned on at half of the overall number of epochs. After turning on causal loss, all parameters are set to untrainable except for the concept vectors, which ensures that the reconstruction ability is not forgotten.

The same hyperparameters are set for the conceptSHAP models, which are also found to generate the optimal performances. The threshold is set to be $\beta = 0.3$, as recommended by the original paper on NLP datasets.

The causal loss regularizer λ_c is set depending on the level of confounding in the dataset. $\lambda_c = 1$ is set for all experiments, except for $\lambda_c = 3$ in the case of IMDB with BERT. A higher λ_c will usually lead to a higher output change (I($\mathcal{C}$) and ΔAcc), accompanied by a decrease in faithfulness (RAcc).

To reproduce, all experiments were run with a random seed of 0. All experiments are conducted on the penultimate layer. The hyperparameters are chosen as an optimal default through grid search. To make the comparison fair, all methods use 10 dimensions to encode. During training, perturbation is performed on the concept that is the most similar to the input.

Dataset	β-TCVAE	kmeans	PCA	conceptSHAP	*HI-concept*
IMDB	475.9	37.7	0.8	199.3	227.2
AG	1525.6	15.51	2.5	1749.65	2242.1

TABLE 3.3: A summary of runtime (in seconds) on datasets for BERT.

3.4.4 Training details

In practice, we only turn on the causal loss after a certain number of epochs (usually half of the overall number of epochs) to make sure that the surrogate model first learns to faithfully reconstruct from the set of concepts before optimizing for the impactful ones. This is because learning the two conflicting objectives at once will usually result in low accuracy. We also note that some contextual information is still needed to maximize the accurate reconstruction of hidden activations $\phi(\mathbf{x})$. Thus, the causality loss is enforced on all concepts except the last one $\mathbf{c}_n$, which is used as a 'context concept'. During model inference, the last (non-impactful) concept is unused.

After training, we post-process discovered concepts to filter out unused ones. While the number of concepts n is user-selected, as in many topic models, it is an inherent flaw as it requires a certain level of domain expertise. For example, in a movie review dataset with only 2 output classes, if an unfamiliar user sets n to 200, the model will naturally discover many noisy concepts and only a few useful ones. To ensure that the noisy concepts are eliminated, we post-process the concepts and filter out the unused ones (with an impact $I_{\text{ind}}(\mathbf{c}_i)$ close to 0). Thus, a more desirable number of concepts is returned even if the user provides an overestimate of n. In our experiments, we see that, after filtering, the model always achieves a better or the same prediction-reconstruction performance as before. However, even with this post-processing, specifying too large a number of concepts can still be dangerous as it harms the concept model's training process.

3.4.5 Run-time

As our model additionally optimizes for causality loss, the run-time is slightly longer than the baseline method ConceptSHAP [18], but remains acceptably efficient for

practical applications. A summary of the run-time is shown in §3.4.5. All models shown are run on the GTX 1080Ti graphic card with 12 GB memory.

The computational overhead introduced by our causality optimization is primarily attributed to the additional gradient calculations required during training. However, this increase in computational cost is justified by the significant improvement in explanation quality and causal fidelity. For instance, our approach adds only a marginal 15-20% increase in training time compared to ConceptSHAP while delivering substantially more reliable causal explanations.

Generally, as post-hoc explainability methods, the runtimes are very light and, therefore, a concern that is less important than the model quality. For example, on a dataset of size 50k such as IMDB, it only takes 227.2 seconds (3.8 minutes) to train our *HI-concept* model. This efficiency makes our approach viable even for larger production models where explanation generation should not significantly impact overall system performance. Furthermore, once trained, the inference time for generating explanations remains comparable to other state-of-the-art methods, making it suitable for real-time or near-real-time explanation requirements in practical applications.

3.5 Results and Analysis

p_{cor}	Cls.Acc	Method	Acc	CACE	ΔAcc
0.50	95.4%	ConceptSHAP	97.6%	0.070	6.1%
		HI-concept	**98.4%**	**0.102**	**9.4%** (+3.3%)
0.65	99.0%	ConceptSHAP	**99.7%**	0.038	3.5%
		HI-concept	99.3%	**0.084**	**6.8%** (+3.4%))
0.75	96.1%	ConceptSHAP	98.3%	0.069	6.0%
		HI-concept	**98.9%**	**0.123**	**12.2%** (+6.2%)

TABLE 3.4: Faithfulness (Acc) and Causality (CACE, ΔAcc) evaluation on the toy dataset. Cls.Acc denotes the original classification model's accuracy.

Dataset	Model	Method	Faithfulness					Causality	
			Acc	Precision	Recall	F1	Completeness	CACE	ΔAcc
IMDB	Transformer	β-TCVAE [94]	43.53%	50.23	50.03	33.08	27.36	0.037	1.50%
		K-means [97]	83.64%	84.74	85.05	83.63	61.87	0.047	2.59%
		PCA [96]	85.18%	85.56	86.20	85.15	**62.36**	0.001	0.01%
		ConceptSHAP [18]	84.36%	85.04	85.56	84.34	62.05	0.031	1.30%
		HI-concept	**88.78%**	**90.07**	**87.50**	**88.24**	58.10	**0.150**	**11.06%**
	BERT	β-TCVAE [94]	93.86%	94.31	93.43	93.68	10.71	0.057	4.05%
		K-means [97]	**98.69%**	96.16	96.23	96.19	15.69	0.037	0.97%
		PCA [96]	96.68%	**96.65**	**96.68**	**96.67**	15.33	0.002	0.02%
		ConceptSHAP [18]	95.84%	95.78	95.96	95.83	17.16	0.050	0.06%
		HI-concept	92.97%	93.25	93.34	92.97	**21.04**	**0.099**	**8.99%**
	T5	β-TCVAE [94]	0.00%	0.00	0.00	0.00	-23.70	0.000	0.00%
		K-means [97]	75.85%	37.92	50.00	43.13	26.83	0.025	1.06
		PCA [96]	98.86%	99.04	97.85	98.43	**48.42**	0.000	0.02%
		ConceptSHAP [18]	0.00%	0.00	0.00	0.00	-23.70	0.000	20.21%
		HI-concept	**99.50%**	**99.65**	**98.98**	**99.31**	**48.87**	**0.153**	**62.47%**
	Llama	β-TCVAE [94]	20.56%	33.41	33.36	13.30	-14.29	0.001	0.15%
		K-means [97]	15.31%	5.10	33.33	8.85	-21.82	0.019	0.00%
		PCA [96]	**95.15%**	**67.97**	**77.66**	**69.80**	**64.19**	0.001	0.03%
		ConceptSHAP [18]	18.83%	42.83	34.95	14.88	-1.78	0.005	1.60%
		HI-concept	87.87%	53.27	68.60	55.29	59.83	**0.042**	**28.69%**
AG-News	Transformer	β-TCVAE [94]	98.91%	98.94	98.94	98.93	**66.73**	0.049	6.62%
		K-means [97]	98.16%	98.32	98.11	98.18	65.99	0.044	0.07%
		PCA [96]	**99.99%**	**99.99**	**99.99**	**99.99**	66.66	0.000	0.03%
		ConceptSHAP [18]	73.01%	59.36	74.34	64.88	47.07	0.000	0.00%
		HI-concept	99.50%	99.50	99.51	99.50	66.70	**0.046**	**7.12%**
	BERT	β-TCVAE [94]	92.30%	94.93	91.89	92.91	57.25	0.044	5.32%
		K-means [97]	86.83%	92.74	85.42	87.53	52.62	0.028	7.15%
		PCA [96]	99.79%	99.82	99.77	99.79	61.04	0.001	0.01%
		ConceptSHAP [18]	93.46%	93.70	94.62	93.66	**62.69**	0.025	4.44%
		HI-concept	**99.90%**	**99.89**	**99.90**	**99.89**	61.12	**0.058**	**10.54%**
	T5	β-TCVAE [94]	0.00%	0.00	0.00	0.00	-20.60	0.000	0.00%
		K-means [97]	24.87%	6.22	25.00	9.96	4.40	0.011	1.49%
		PCA [96]	97.38%	97.40	97.37	97.38	73.12	0.000	0.01%
		ConceptSHAP [18]	0.00%	0.00	0.00	0.00	-20.60	0.000	0.01%
		HI-concept	**99.46%**	**99.46**	**99.46**	**99.46**	**73.70**	**0.075**	**72.37%**
	Llama	β-TCVAE [94]	1.27%	0.25	20.00	0.50	-23.89	0.000	0.01%
		K-means [97]	37.00%	7.40	20.00	10.80	1.09	0.007	0.02%
		PCA [96]	**85.41%**	**65.78**	**67.98**	**66.73**	**51.46**	0.000	0.03%
		ConceptSHAP [18]	17.01%	35.37	35.20	15.87	-7.73	0.002	2.83%
		HI-concept	81.52%	48.59	55.99	51.53	43.07	**0.039**	**27.79%**

TABLE 3.5: Faithfulness (Acc, Precision, Recall, F1, Completeness) and causality (CACE, ΔAcc) evaluation of different text classification methods. The best result is bolded, and the second-best result is underlined.

3.5.1 Sanity Check

To first provide a sanity check for our method, we follow the toy experiment design in Yeh et al. [18], which explains a CNN model trained on a synthetic graphic dataset. To mimic the confounding effects ($X \rightarrow E$) as in Fig. 3.3, we add correlations (controlled by p_{cor}) among ground truth concepts. Then, we compared the concepts discovered by *HI-concept* with ConceptSHAP.

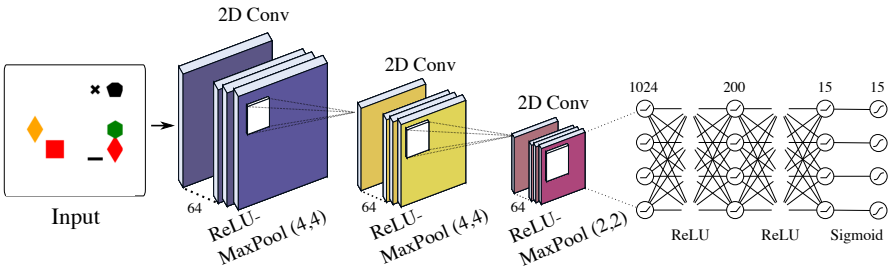


FIGURE 3.4: Convolutional Neural Network used for classifying the toy dataset.

We conduct experiments on a synthetic (toy) image dataset with ground truth concepts in order to test the validity of our method and confirm the claim that higher confounding effects within the dataset lead to more correlational explanations, thus calling for a more causal explainability approach. Specifically, we extend the toy dataset design of Yeh et al. [18] to make it more realistic by inserting spurious correlations.

Data generation. As a synthetic setup, at most 15 shapes are randomly scattered on a blank canvas at random locations with random color selections (as noise). For each image sample x_j, $\boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{\{1:15\}}}$ are binary variables of whether or not a shape is present in x_j with each $\boldsymbol{z}^j_s$ sampling from a Bernoulli distribution with probability 0.5. Then, a 15-class target $\boldsymbol{y_j}$ is constructed with respect to whether the first 5 shapes ($\boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{\{1:5\}}}$) are present or not with human-designed rules. For example, $\boldsymbol{y_1} = \sim (\boldsymbol{z_1} \cdot \boldsymbol{z_3}) + \boldsymbol{z_4}$. A total of $60,000$ examples are generated as the toy dataset using a seed of 0.

The setup mentioned above is, in fact, far away from realistic scenarios, as it does not consider possible confounding. Thus, to make it more realistic, we insert spurious correlations between the pairs $(\boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{\{1:5\}}}, \boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{\{6:10\}}}), (\boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{\{6:10\}}}, \boldsymbol{z}^{\boldsymbol{j}}_{\boldsymbol{\{11:15\}}})$ with a correlation factor p_{cor}. For example, when $z_1 = 1$, $z_6 = \text{Bernoulli}(p_{\text{cor}})$; when $z_1 = 0$, $z_6 = \text{Bernoulli}(1 - p_{\text{cor}})$.

CNN classification model used for the toy example. The CNN classification model used for the toy dataset is shown in Fig. 3.4. Specifically, 3 convolutional layers with a kernel size of 5 and 64 output channels were used, each followed by a ReLU activation and max pooling layer. Then, the result is flattened into a linear vector, followed by 2 linear layers and a sigmoid activation function. The output is a 15-dimensional binary classification probability. The model is trained for 100

Method	CACE	Keywords
CS	0.134	apple, NASA, Microsoft, new, sun, red, super, game
CS	0.000	one, two, gt, new, cl, lt, first, world, mo, last
HI-C	0.130	us, bush, u, eu, new, peoples, china, high, gt, world
HI-C	0.003	us, update, new, mo, two, first, knicks, last, one, hen

TABLE 3.6: Generated concepts with Average Impact (CACE) from AG-News dataset, BERT model. CS is ConceptSHAP, HI-C is *HI-concept*. Each line is one concept, represented by keywords, which are ordered by descending importance.

epochs with an Adam optimizer with learning rate $3e-4$. For reproducibility purposes, the model is initialized and trained with a seed of 0.

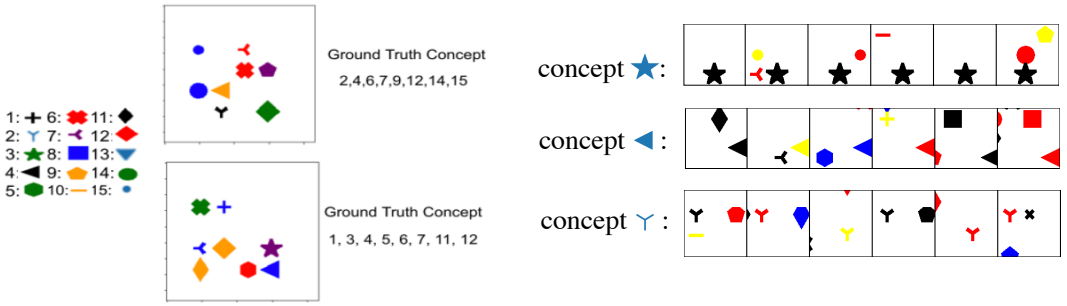


FIGURE 3.5: Examples from the toy dataset and concepts discovered.

Visualizations. As an example visualization, in Fig. 3.5, two random images from the toy dataset are displayed on the left, while three example concepts discovered by *HI-concept* are plotted on the right. We could observe that *HI-concept* is able to derive meaningful clusters as concepts, which provide a sanity check for usability of the latent concepts.

Results on toy dataset. From the results shown in Table 3.4, we could observe that, as we increase p_{cor} to mimic an increase in confounding levels in real life, our method discovers concepts that consistently outperforms the baseline by deriving more impactful features. As confounding levels (p_{cor}) in the dataset increase, the performance gap (ΔAcc) also widens. At the same time, HI-Concept maintains the best RAcc, indicating faithfulness to the original predictions. Therefore, *HI-concept* successfully improves explanatory impact, especially for highly correlational tasks and datasets. Moreover, we note that the improvement is even stronger in real data experiments, as the added artificial confounding is more complicated in real-life scenarios.

3.5.2 Quantitative Results on Text Classification

The results of the experiment on the text classification datasets are presented in Table 3.5. Overall, HI-Concept not only achieves the best performance in causality, but improves on faithfulness as well. For faithfulness metrics (Acc, Precision, Recall, F1, and Completeness), *HI-concept* achieves the best or second-best results for almost all datasets and models. Notably, for the cases achieving second-best performance, the best model for faithfulness is PCA. PCA, however, as a completely different group of methods, is often faced with the issue of low causal impact (shows CACE close to 0 in Table 2). While considering causality metrics (CACE and ΔAcc), our *HI-concept* exhibits a significantly greater superiority. Causality metrics for baseline methods are mostly minimal, which implies that most explanatory features discovered are correlational and unreliable. In comparison, the concepts discovered by *HI-concept* show significant improvements in both causality and faithfulness, especially for pretrained models such as BERT, Llama, and T5.

This validates the hypothesis that HI-Concept can result in more improvements for larger pre-trained models with more complex architectures. With more parameters and pretraining, these models could encode more correlational information and contain more spurious correlations. As shown with the toy example in §3.5.1, HI-Concept's causality awareness would be more beneficial in highly correlational scenarios.

3.5.3 Qualitative Analysis of Text Classification

We take a closer look at BERT for AG-News to qualitatively examine the concepts discovered in terms of *causality* and *explainability*.

Causality. Table 3.6 visualizes the most and least causal concepts obtained from both baseline ConceptSHAP and our *HI-concept*. The words are organized by descending concept importance scores (described in §3.3.3). For the most causal concept (*i.e.*,, larger CACE), the one by ConceptSHAP implies technological news, but has some confounding keywords from the sports category (*e.g.*,, "red", "super", "game"). The one by *HI-concept* clearly points to political news, without confounding words that belong to other categories. While for the least causal concept, the

Method	Visualization
ConceptSHAP	dream team leads spain 44 - 42 at halftime athens, greece - as expected, the u.s. men's basketball team had its hands full in a quarterfinal game against spain on thursday...
HI-concept	dream team leads spain 44 - 42 at halftime athens, greece - as expected, the u.s. men's basketball team had its hands full in a quarterfinal game against spain on thursday ...

FIGURE 3.6: Qualitative comparison from AG-News: "World" news misclassified as "Sports" by BERT.

	Accuracy	Confidence	Time Spent
Plain	72.5%	3.2	10.7
ConceptSHAP	68.5%	2.7	10.6
Polyjuice	73.5%	2.6	7.6
HI-concept	**80.5%**	**3.5**	**9.3**

TABLE 3.7: Human study for explainability evaluation.

ConceptSHAP only consists of purely correlational and non-semantically meaningful words. Instead, *HI-concept* still contains class-specific words (*e.g.*,, "us", "knicks"), which result in non-zero CACE. Overall, *HI-concept* results in a set of more task-relevant and semantically meaningful concepts.

Explainability. Fig. 3.6 shows the failure case ("World" news misclassified as "Sports") highlighted with the top concept discovered. ConceptSHAP discovers a top concept related to the keywords "leads", "as expected", or "on thursday", which are not informative as to why the model classified this input as "Sports". On the contrary, *HI-concept* could precisely point out why: BERT is looking at keywords such as "dream team", "game", and country names. Such examples show the potential of *HI-concept* being used in understanding the model's failure processes, which we further investigate in §3.5.5 with a carefully designed human study.

3.5.4 Generalization to Concept Insertion

As mentioned in §3.3.4, we study the causal impact of concepts by generalizing to a

novel *insertion* operation. With the insertion of the found concepts one by one, we expect to observe *gradual improvement* of the recovering accuracy of the concept model. For example, we first evaluate the concept model (with 10 concepts) with only the most important concept, while masking all other concepts. Then, we evaluate the concept model with the two most important concepts, while masking all other concepts. The process goes on until we mask 0 concepts. As we unmask more and more concepts, the model performance is expected to gradually improve in order for each concept to have some causal importance. Formally, at the step $m \in 1, \ldots, n$, the concept model reconstruction becomes $g_\theta(p_c(x_j)|c_{i \in C \setminus C_m} = 0)$, where C_m is the set of most important m concepts.

Fig. 3.11 shows the trend results on the AG-News dataset. The concept is inserted in the order of descending importance. Obviously, our *HI-concept*, plotted as the red line, is the only method that shows gradual improvement consistently for all base models. While for other comparison methods, a single concept can already result in maximum accuracy, *e.g.,*, all baselines on T5 and Llama, indicating less-causal sets of concepts overall.

3.5.5 Human Study

Design and Baselines. To systematically test whether derived features are explainable to humans, we design a human study to test the degree to which "a user can correctly and efficiently predict the method's results", which is the explainability definition by Kim et al. [6]. Inspired by the forward simulation design from Hase and Bansal [98], we carefully conduct the following human study: We first show 100 randomly selected examples from AG's test set to users and ask them to predict the model's news topic classification. Then, we show the same examples again but with assistive information from *HI-concept*, including textual highlights and topic keywords, and ask users to predict the model's decision again. As a comparison, we show examples augmented by ConceptSHAP instead. For each question, we let users rate their confidence and record the time spent in seconds. Moreover, to test against local counterfactuals, which is a popular group of explainability methods, we also include Polyjuice [99] as another baseline. Polyjuice is a generator method that utilizes a finetuned GPT-2 model for producing diverse local counterfactuals to a sentence. Thus, it enables an automated approach to

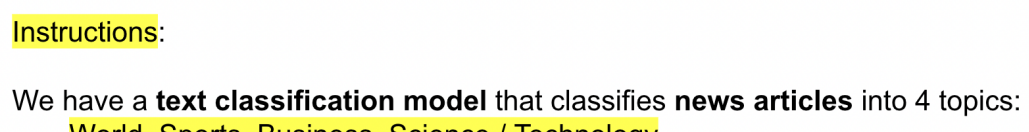


FIGURE 3.7: Human study instructions for plain examples.

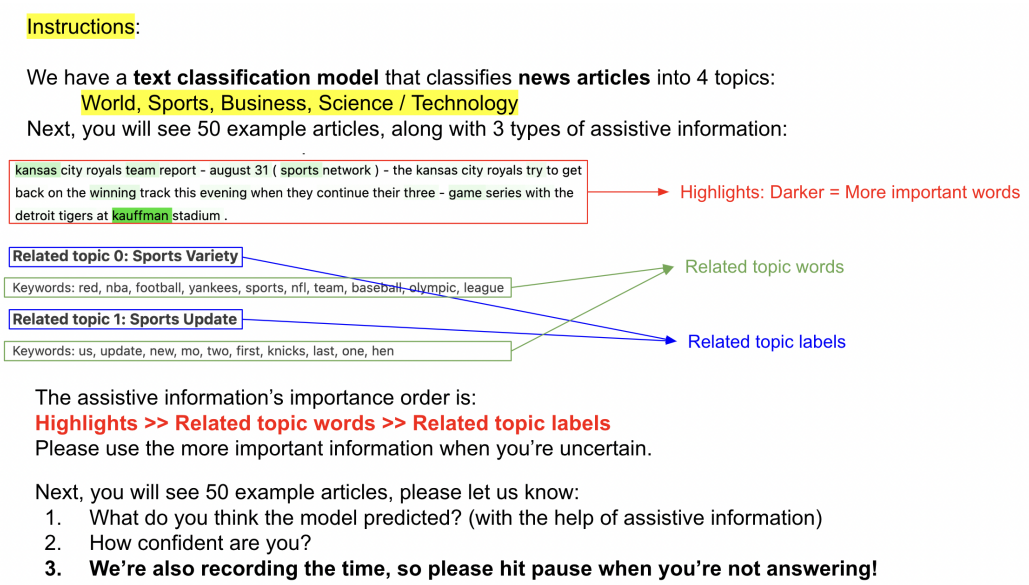


FIGURE 3.8: Human study instructions for *HI-concept* augmented examples.

derive token explanations with Shapley values. Ideally, good explanations could help users better predict the model outcomes, thus increasing usability by resulting in higher accuracy and higher confidence.

Details for Conducting Human Study. For the human study, 100 examples are randomly selected from the test set $\mathcal{D}_{\text{test}}$. The questionnaire takes the format of a self-constructed website. Firstly, we show the examples without any assistive information, where the instructions are shown in Fig. 3.7 and an example question looks like Fig. 3.9. Secondly, the same examples are shown with assistive information derived from ConceptSHAP. Lastly, the examples are shown with assistive information derived from *HI-Concept*. The instructions are shown in Fig. 3.8 and an example question looks like Fig. 3.10. 4 volunteers (Ph.D. students) each answered 50 plain examples and 50 augmented examples. The volunteers are all proficient

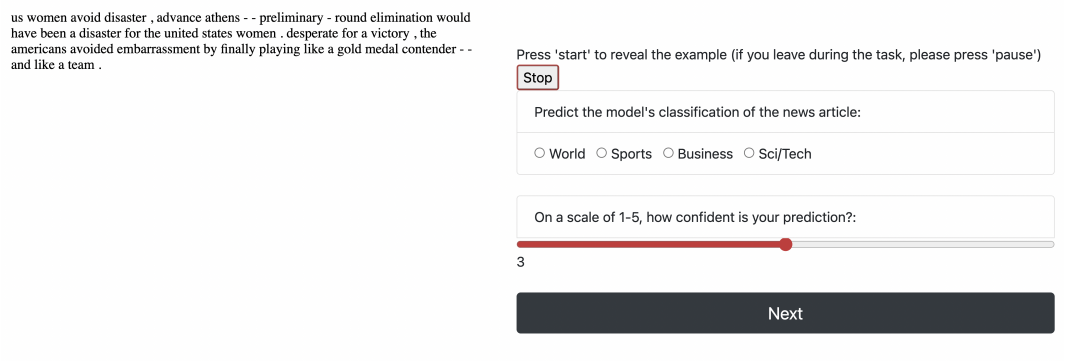


FIGURE 3.9: Human study question and answer.

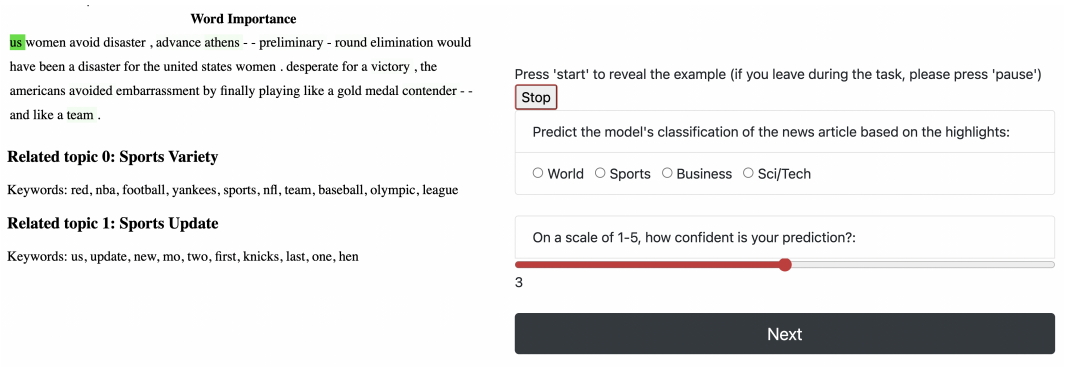


FIGURE 3.10: Human study question and answer.

in English. The volunteers report an average time of approximately 30 minutes for answering all 100 questions. As the volunteers are working also in AI-related areas and are briefed about the purpose and usage of survey data beforehand, they understand fully the data collection and usage. Thus, implicit consent is granted by participation.

As one resulting concept is "a group of words that are meaningful" [82], which could take some time for humans to read, we also employ an LLM (GPT-3.5) to summarize the words into an assistive label. The resulting labels allow humans to quickly grasp the gist of an abstract concept. Specifically, we used the GPT-3.5-turbo model with the following prompt:

"You're an expert in topic labeling. Please come up with a short word or phrase that summarizes the topic with the keywords below:

[set of keywords]"

Human Study Results. As shown in Table 3.7, when the users are given assistive information provided by *HI-concept*, their accuracy of predicting the model's

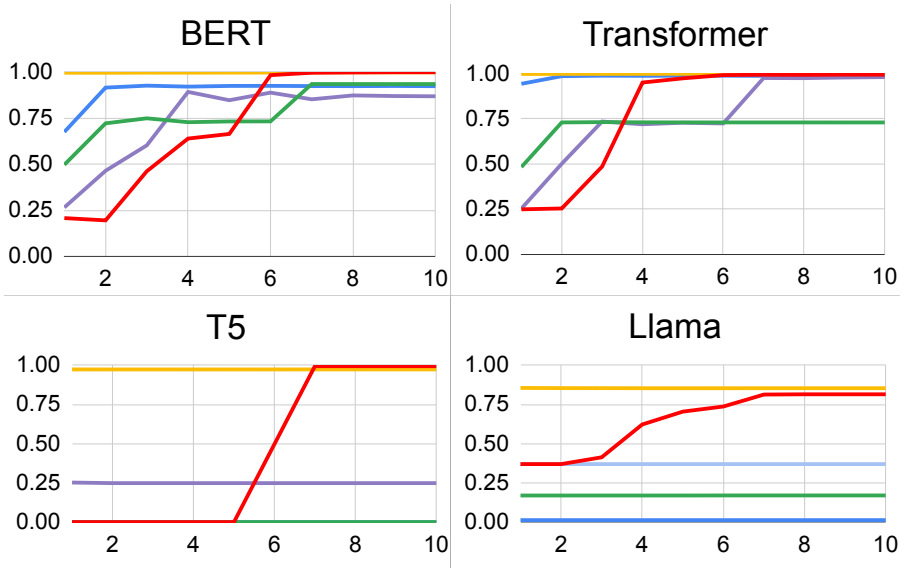


FIGURE 3.11: Effects of concept insertion on accuracy on AG-News dataset. Each figure represents a different model where the number of inserted concepts (x-axis) is plotted against accuracy (y-axis).

decisions improved from 72.5% to 80.5%. On average, users also report higher confidence in their predictions and spend less time on the questions. When given correlational explanations by ConceptSHAP, however, both prediction accuracy and confidence decrease. Polyjuice, as a local counterfactual baseline, results in a human prediction accuracy of 73.5%. It surpasses the conceptSHAP baseline (68.5%) but still lags behind HI-Concept (80.5%). Moreover, HI-Concept also maintains the highest confidence score over all the baselines, outperforming Polyjuice by 1.1 (on a scale of 1-5). We note that users with Polyjuice spend less time (7.6s V.S. 9.3s of HI-Concept) for the decision. It could be because Polyjuice tends to assign high importance to a selected few words, while giving minimal importance to others. This leads to quicker decision-making by users but is also accompanied by low accuracy and confidence. Overall, our study achieves the Cohen's Kappa agreement of 0.74, which is considered substantial agreement [100].

3.5.6 Ablation Study

To further investigate the effect of different loss objectives and various hyperparameters, we conduct multiple ablation studies.

Method	Acc	CACE	ΔAcc
Without Auto-Encoding Loss	93.46%	0.028	6.11%
Without Prediction Loss	68.00%	0.035	**17.41%**
Without Regularizer Loss	95.76%	0.041	6.23%
Without Causality Loss	**99.92%**	0.029	2.95%
HI-concept	99.90%	**0.058**	10.54%

TABLE 3.8: Ablation on BERT for IMDB with faithfulness (Acc) and impact (CACE, ΔAcc) evaluation.

Loss objectives. To ensure that the designated 4 objectives behave as expected, we conduct ablation studies for BERT on AG-News and report the results in Table 3.8. As observed, each designed loss plays its own role. Specifically, eliminating prediction loss leads to a large decrease in Acc, resulting in an unfaithful model. Therefore, even though its model explanations are more causal (large ΔAcc), the results cannot be trusted. Meanwhile, the auto-encoding and regularizer loss contribute to both faithfulness and causality, while causality loss mostly helps to ensure the causal metric. The full *HI-concept* method discovers a set of concepts with both good causality and faithfulness.

Hyperparameters. The proposed method of *HI-concept* includes many tunable hyperparameters, including the top-N neighborhood, threshold, etc. While these parameters are set at the default mentioned in §3.4.3, there are two hyperparameters that users can customize the most: the layer to interpret at and the number of concepts. To better understand how these two parameters may affect the generated concepts, we conduct comparisons on both. We evaluate in terms of impact and topic quality. For impact, we have reported the number of effective concepts left after post-processing, the recovering accuracy (RAcc), the Average Impact (I($\mathcal{C}$)), and the induced change in accuracy (ΔAcc). For topic quality, we have reported coherence scores, including averaged Pointwise Mutual Information (PMI) (c_uci score), normalized PMI (c_npmi score), c_v score which measures how often the topic words appear together in the corpus, and word2vec similarity [101].

The following comparisons are all conducted on the AG-news dataset with BERT, where the other hyperparameters mentioned in §3.4.3 stay the same.

(i) Layer to Interpret. To compare what each layer discovered, as BERT has 12 layers, we experimented on the 3rd, 6th, 9th, and penultimate layer respectively,

all with 10 concepts. Overall, the later layers tend to discover more class-coherent concepts. The beginning layers, however, could discover more abstract features and also lexical word clusters, such as concepts with only nouns or adjectives. This finding is confirmed by a study of topic coherence metrics and findings from Dalvi et al. [82], where they observe that BERT finds more lexical information in the earlier layers.

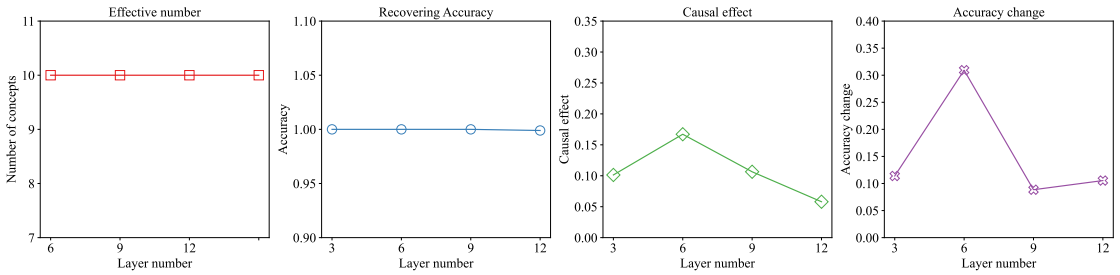


FIGURE 3.12: Layer-wise effective number of concepts, RAcc ↑, I($\mathcal{C}$) ↑, and Δ Acc ↑.

Quantitatively, we plotted out the effective number of concepts, recovering accuracy, impact, and accuracy change in Fig. 3.12. All layers demonstrate similar performances in recovering accuracy, which is close to 100%. The intermediate layers, especially the 6th layer, produce a higher average impact and recovering accuracy. This is because the intermediate layers discover concepts on the token-level, while the penultimate layer concepts are sentence-level (on the [CLS] token). Thus, the token-level concepts will have more fine-grained control.

Qualitatively, we plotted some wordclouds of the keywords in discovered concepts in Fig. 3.13. We could see that, in the penultimate layer, concepts are more concentrated on each class. For example, the first concept would correspond to the class "Sports", the second to "Sci/Tech", and the third to "World" news. The emphasis on events is also clearer, such as the third one talking about the Iraq War. However, When we move to earlier layers, the concepts' class labels are more mixed together. In the 9th layer, the first concept concerns government, which includes terms such as "government", "internet", "security", "bomb", "baseball", etc. It could, however, correspond to many class labels, such as "Sci/Tech", "World", or even "Sports". Similarity, the second concept talks about China, including "china", "billion", "people", "activitists", "announcement", etc. The third concept is interesting as it covers mostly adjective words which do not seem to correlate too much in semantic meanings, such as "low", "big", "closer", and "third". Similar

(A) Layer 9 (B) Layer 12

FIGURE 3.13: Wordclouds of concepts generated on the 9th (left) and 12th (right) layer. The 9th layer includes a government concept, a China concept, and an Adjective (mostly) concept. The 12th layer includes a sports concept, a technology concept, and a political concept.

observations are also confirmed in papers such as [82], which derives concepts using agglomerative hierarchical clustering combined with human annotations in BERT latent representations. They observe that BERT finds more lexical information in the earlier layers.

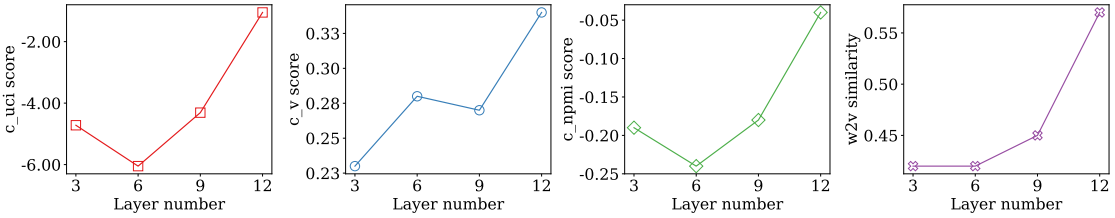


FIGURE 3.14: Layer-wise Topic Coherence Comparison.

In terms of topic quality, we evaluated the concept keywords using coherence metrics. As shown in Fig. 3.14, all coherence scores showed a general trend of concepts

becoming more coherent as the layer number increases. The conclusion is consistent with the wordcloud visualizations.

Thus, in real-life debugging scenarios, we recommend using the penultimate layer, which will find more coherent topics. However, there could be continued work to discover information learned in the prior layers and to investigate how information flows through layers in a hierarchical way.

(ii) Number of Concepts. We experiment with 3, 5, 10, 50, and 100 concepts on the penultimate layer. We find that a concept number close to the number of output classes usually gives higher prediction changes, while increasing the number results in higher recovering accuracy. When the number of concepts becomes larger, concepts usually become more coherent. However, with too large a number of concepts, the performance will decrease, as more noise is introduced into the training process.

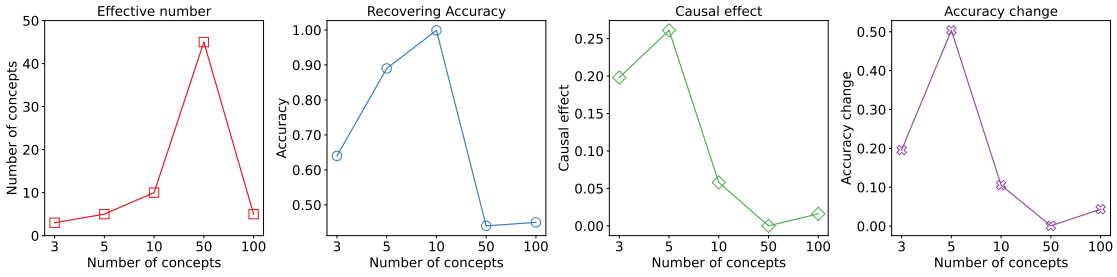


FIGURE 3.15: Concept-wise effective number of concepts, RAcc ↑, I($\mathcal{C}$) ↑, and Δ Acc ↑.

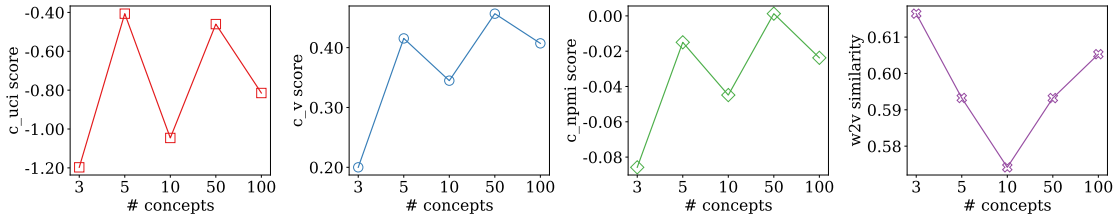


FIGURE 3.16: Concept-wise Topic Coherence Comparison.

From Fig. 3.15, we could see that the performance is very dependent on the number of concepts. The effective number of concepts, recovering accuracy, average impact, and accuracy change all appear to be elbow-shaped. In this case, 5 concepts provided the highest impact on output predictions, as it is close to the number of classes (4) in the AG-News dataset. Increasing the number of concepts to 10 would yield a better recovering accuracy. As the number of concepts increases to

50 and 100, we observe that the model fails to learn completely. In practice, we have often observed the best number to be positively correlated with the number of dataset classes. In other words, a dataset with more classes will require a higher number of concepts for faithful reconstruction. In terms of topic coherence, we could observe from Fig. 3.16 that the topic coherence scores usually oscillate, but mostly display a generally upward trend of becoming more coherent as the number of concepts increases.

3.6 Conclusions

We propose *HI-concept* to derive impactful concepts to explain the black-box language model's decisions. Our framework not only derives high-impact concepts that mitigate the confounding issue with the proposed causal objective, but also advances previous evaluations via both quantitative global accuracy change and qualitative insertion study. Extensive experiments, visualizations, figures, and human studies prove that our *HI-concept* can produce semantically coherent and user-friendly concept explanations.

HI-concept demonstrates the potential to play an important role in practical scenarios such as debugging and transparency. As AI ethics have become a major concern in real-life applications, such explanations can help users better identify bias and promote fairness. As a future venue to our work, we believe that our framework will set a good foundation for future research on causal NLP explainability methods, especially those that hope to derive human-friendly explanations. As for potential concerns, *HI-concept* only encourages high impact in post-hoc model explanations and should serve as an assistive tool instead of being accepted as ground-truth. Thus, to improve it further, a similar causal objective could be used to address spurious correlations during training. It also has the potential of being carried over to other domains, such as vision or tabular tasks. The high-level attributes in the hidden space can also be used in downstream applications to provide better controllability for the users.

Chapter 4

Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework

As large language models (LLMs) have become the norm in NLP, demonstrating good performance in generation and reasoning tasks, one of its most fatal disadvantages is the lack of factual correctness. Generating unfactual texts not only leads to lower performances but also degrades the trust and validity of their applications. Chain-of-Thought (CoT) prompting improves trust and model performance on complex reasoning tasks by generating interpretable reasoning chains, but still suffers from factuality concerns in knowledge-intensive tasks. In this paper, we propose the Verify-and-Edit framework for CoT prompting, which seeks to increase prediction factuality by post-editing reasoning chains according to external knowledge. Building on top of GPT-3, our framework leads to accuracy improvements in multiple open-domain question-answering tasks. For reproducing our results and extending the framework further, we make our codebase available at https://github.com/RuochenZhao/Verify-and-Edit.

4.1 Chapter Background

Large Language Models (LLMs) have become the new norm in many downstream NLP tasks. In utilizing these LLMs, Chain-of-Thought (CoT) prompting [1] is

found to improve performances for tasks that require complex reasoning, such as math word problems, commonsense reasoning, and symbolic manipulation. At the same time, it is able to generate interpretable reasoning chains. Recent work further explored how to use these reasoning chains to select better predictions. However, the primary focus of these methods has been to improve end-task performance by utilizing generated CoTs as-is. For example, Ye and Durrett [48] trains a calibrator that tunes prediction probabilities based on rationale scores; Wang et al. [2] samples multiple reasoning paths to find the most common (consistent) prediction. Only a few, such as Creswell et al. [64] and Zhou et al. [63], have explored ways to improve the quality of CoTs themselves.

In fact, improving the CoT quality could be beneficial in enhancing both interpretability and end-task performance. Ye and Durrett [48] point out that explanations judged as good by humans often indicate more accurate predictions. Intuitively, a better set of CoT prompts could provide better grounding and logically consistent thought processes, thus leading to more accurate predictions.

To improve generation quality, one important aspect is *factual correctness*, which is currently one of the most fatal drawbacks of LLMs [102, 103]. In answering user queries, LLMs such as GPT-3 [71] tend to make up facts and details, which is now flagged as a primary warning in their API usage. As a major use case of LLMs is the prospect of replacing traditional search engines and usage for more direct information access through question-answering, factuality concerns could largely undermine their validity and degrade users' level of trust [104]. Fixing this issue is challenging and the concerns still persist even after the models are instruction-tuned with human feedback [74]. This is because the source of truth can be unavailable during the finetuning process [102].

Thus, it is of urgent concern to better control the generation and increase the factual correctness of predictions. As LLMs could fail to recall accurate details when functioning as a knowledge base [48, 64], if possible, knowledge from external sources could be introduced as assistance. Assisted thought process is also common in human reasoning: when humans answer questions, they often search (or revisit) external knowledge sources for supporting facts in order to refresh their (internal) memory.

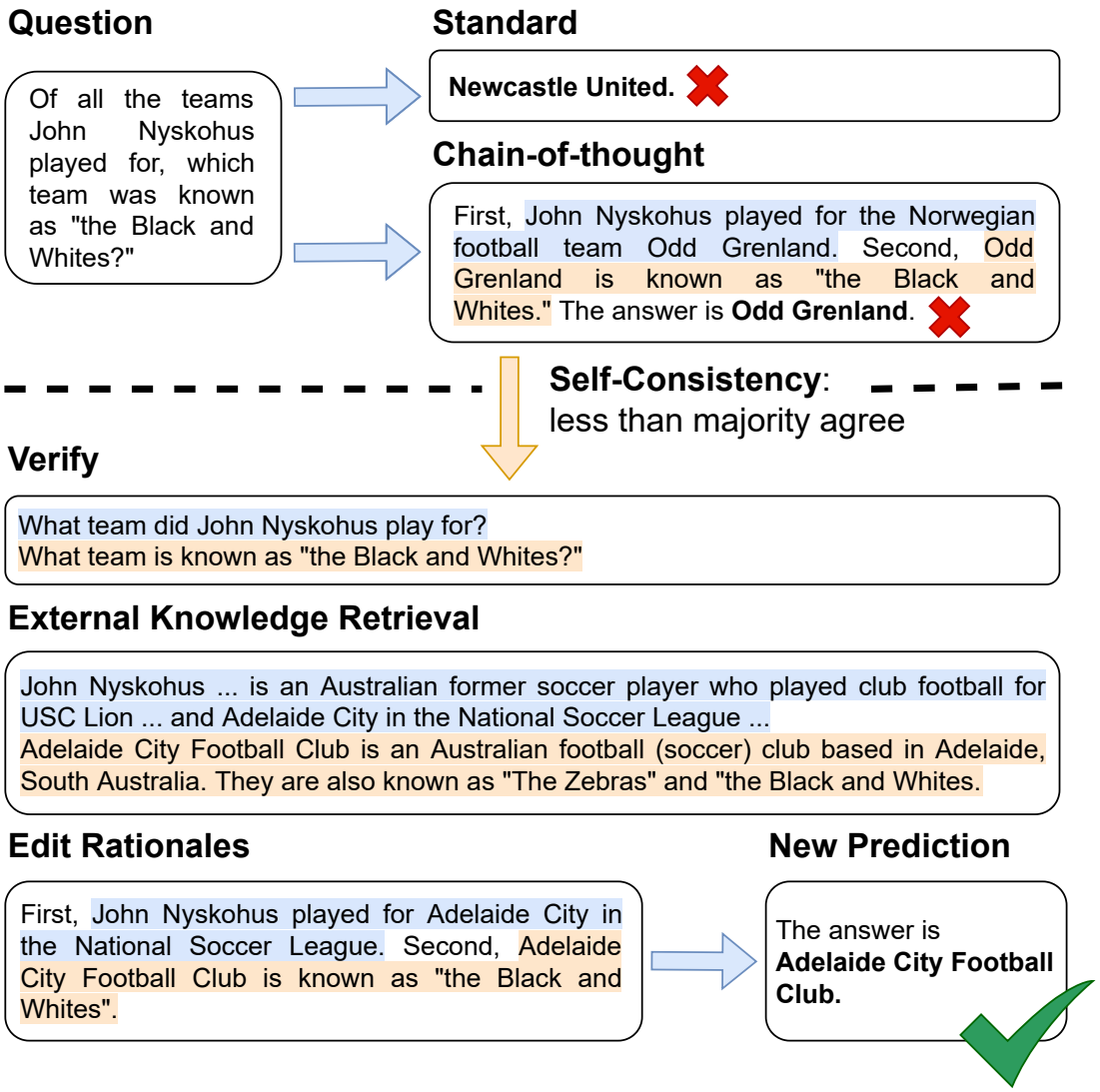


FIGURE 4.1: The Verify-and-Edit framework consists of five steps: (1) pass predictions with lower-than-average consistency to the next stages while leaving highly consistent predictions as-is; (2) produce verifying questions; (3) retrieve external knowledge; (4) edit rationales with informed answers; and (5) produce new predictions.

Inspired by this, in this work we propose a **Verify-and-Edit** (VE) framework to post-edit the reasoning chains for more factually aligned predictions. As shown in Fig. 4.1, we first select uncertain instances to edit, which have a less-than-majority-agree consistency. These instances, as implied by Wang et al. [2], often consist of plausible-sounding statements, such as the sentence "John Nyskohus played for the Norweigian football team Odd Greenland" in Fig. 4.1. When editing, we first generate a question to verify this detail, such as "What team did John Nyskohus play for?" Then, to answer this query, we introduce external knowledge through

open-domain retrieval systems. For example, the fact "John Nyskohus ... played for Adelaide City.." is retrieved in this instance. Then, the rationales are edited by providing the retrieved facts in the prompts as memory refreshments. Thus, the edited rationales could be updated corresponding to the retrieved facts (Fig. 4.1). Given the edited rationales, the new prediction is generated, which considers more factually aligned reasoning traces.

To our knowledge, our work is the first to post-edit CoT-style reasoning chains to enhance prediction performance. We perform experiments on two open-domain Question Answering (QA) tasks that require reasoning: Adversarial HotpotQA [105] and 2WikiMultihop [106]. We also test its performance on the Fact Verification task using Fever [55]. We find that the model is able to benefit from more factual reasoning chains, thus generating more accurate predictions. For example, for open-domain QA, our model demonstrates 3.8x accuracy improvement compared to similar retrieval-augmented models on AdvHotpot. On 2WikiMultihop, Verify-and-Edit reaches 33.6% accuracy with open-domain search, while CoT Self-Consistency stands at 27.7%. As a post-training method, the framework complements traditional fine-tuning methodologies in further addressing knowledge gaps. Given the appropriate knowledge source, the framework also has the potential of replacing fine-tuning methods with a zero-shot manner.

4.2 Verify-and-Edit Framework

Our goal is to make LLMs generate more factual reasoning chains with CoT prompting assisted with external knowledge, thereby also improving prediction accuracy of the final answer. We hypothesize that this can enhance LLMs' capability to solve complex knowledge-intensive tasks that require multiple reasoning steps to arrive at an answer.

Generally, we hope to follow the human reasoning process: when a person answers a question, if he/she is unsure, he/she would search for a supporting fact and consider it before giving the final answer. Thus, we could separate the Verify-and-Edit (VE) framework into 3 different stages: finding uncertain predictions, editing their rationales by searching for supporting facts, and using the edited rationales to generate final answers (Fig. 4.1). In designing the stages, we hope to maximally

Algorithm 1 Verify-and-Edit

Require: The original question q; An n-shot CoT prompt p_{cot}
Require: An LLM $f(\cdot)$;
LM number of completions n;
LM decoding temperature τ
Require: An external knowledge retrieval model $g(\cdot)$
Require: n-shot prompts for verifying question generation (p_{vq});
n-shot prompts for answer generation (p_{va})
$R, A \leftarrow f(p_{cot}, q, n, \tau)$ ▷ Generate a set of reasonings (R) and answers (A).
$s^*_{sc} \leftarrow \max P(a|p_{cot}, q), a \in A$ ▷ The highest self-consistency score among all answers.
$r^*, a^* \leftarrow \arg\max P(a|p_{cot}, q), a \in A$ ▷ Reasoning and answer with highest self-consistency.
if $s^*_{sc} < \lceil \frac{n}{2} \rceil$ **then** ▷ Edit reasoning with a less-than-majority-agree consistency.
  for $o_i \in r^*$ **do** ▷ Edit each sentence in the reasoning.
    $u \leftarrow f(p_{vq}, q, o_i)$ ▷ Generate verifying question.
    $v \leftarrow g(u)$ ▷ Retrieve external knowledge.
    $w \leftarrow f(p_{va}, u, v)$ ▷ Generate verifying answer.
    $o_i \leftarrow w$ ▷ Edit original reasoning sentence with verifying answer.
  end for
  $a^* \leftarrow f(p_{cot}, q, r^*)$ ▷ Generate final answer with edited reasoning.
  return a^*
else if $s^*_{sc} \geq \lceil \frac{n}{2} \rceil$ **then** ▷ Answer with high consistency is left as-is.
  return a^*
end if

preserve the LLMs' biggest advantage: their open-generation and reasoning ability. And we aim to design tasks and setups as natural and conversational as possible, thus making it easy to understand for humans and LLMs which are trained with natural texts.

4.2.1 Deciding when to edit

How can we identify when a model is unsure of its prediction? The self-consistency method [2] provides a solution. In sampling diverse reasoning paths and answers, self-consistency is found to be highly correlated with accuracy, suggesting that it could provide an uncertainty estimate and confer abilities for the model to "know when it doesn't know". Thus, we begin the VE framework by using the consistency method to sample n diverse reasoning paths for a prediction task. The highly

consistent predictions are left as-is. When consistency is lower than $\lceil n/2 \rceil$, *i.e.,* the majority cannot agree on the same answer, we label it as "uncertain".

4.2.2 How to edit a specific rationale

The rationale, *i.e.,* the thought process (CoT), could be viewed in two parts: facts and reasoning which combines facts to derive a new claim. Thus, we consider improving the CoT from both aspects.

Facts.

To make the thought process more factually correct, we search for supporting facts in external knowledge sources (*e.g.,* Wikipedia).

First, to mimic a human's query when searching for validating facts, a natural question is generated to verify the rationale. For this, we use the in-context learning capability of the same LLM. The original question and the rationale are both provided in the prompt for verifying question generation to ensure that it asks for the most relevant information required to answer the original question, instead of other entities in the rationale. For example, if the rationale (wrong) is "the US president born on 4 August 1961 is John Kennedy." and the original question is "who is the spouse of the US president born on 4 August 1961", we expect the generated verifying question to be: "Who is the US president born on 4 August 1961?" instead of "When is John Kennedy's birthday?" By generating a relevant question instead of directly querying with the generated rationale, we eliminate potential noise brought by incorrect fact generation. In the example above, if one retrieves using the wrong claim "the US president born on 4 August 1961 is John Kennedy", the incorrect entity "John Kennedy" may obfusticate the search process.

In this paper, we use relevant contexts retrieved from 3 systems: (*i*) DrQA [107], an open-domain question-answering system; (*ii*) Wikipedia search of relevant pages.

As the retrieved contexts from a retrieval system could be longer than desired, we use a pre-trained LM to rank and select the top-k sentences most similar to the verifying question query.

Reasoning

While methods such as Selection-Inference [64] directly use retrieved facts as rationales, they are usually too verbose, longer than desired, or contain irrelevant details. Ye and Durrett [48] have made similar observations: directly using supporting sentences is usually too verbose and not sufficient.

To obtain more relevant and logical rationales, we again utilize a natural and generative approach, as reasoning abilities are believed to be already built into LLMs [1]. In particular, by feeding in prompts in the format of "question, rationale, answer", the LLM learns to reason for a few steps before answer generation. Upon investigating the original rationales, we observe that, even when they contain incorrect facts, the logical reasoning component seems to be generally intact. Thus, we use the verifying questions (as logic) and retrieved facts (as information) to generate informed answers. The informed answers are then composed into a new rationale, providing potentially a more factual CoT.

4.2.3 Answering again

Finally, with the post-edited CoT, new answers are generated by prompting the LLM. A pseudocode of the overall procedure is given in Alg. 1, and illustrated with an example in Fig. 4.1 . We can see that, by allowing the LLM to incorporate external knowledge, our method could result in more factually-grounded rationales. When prompted into the LLM as a CoT, it could bring in the information necessary to make a new prediction, which was originally not remembered correctly by the model.

Compared to specifically designed prompts such as ReAct [61], the Verify-and-Edit framework is simple and arguably more natural. Its conversational nature could allow humans to better understand the model's thought processes and have the potential for users to naturally interfere and revise at any stage of inference. In the experiments presented next, we also observe that such a setup is effective in mitigating factuality concerns and boosting end-task performances.

4.3 Experiment Setup

4.3.1 Reasoning tasks

As the Verify-and-Edit framework offers more knowledge-grounded reasoning steps, it should benefit tasks that fulfill the following two properties: (*i*) reliant on multi-hop reasoning to arrive at a later prediction, thus depending on rationale generation, and (*ii*) open-domain, thus needing to interact with an external knowledge source.

Therefore, we validate the approach on three datasets: (*i*) **Adversarial HotpotQA** [105], a multi-hop question answering dataset. We use the challenging subset proposed by Ye and Durrett [48], where the correct and incorrect predictions are balanced using their model. (*ii*) **2WikiMultihop** [106] a multi-hop question-answering dataset exploiting the structured format in Wikidata and use logical rules.[1] (*iii*) **Fever** [55], a fact verification dataset that labels claims as "SUPPORTS", "REFUTES", or "NOT ENOUGH INFO" based on evidence paragraphs from Wikipedia. Similar to the HotpotQA setup, we sample a challenging set by balancing the samples where GPT3 CoT makes correct and incorrect predictions. Specifically, we describe the processing procedures below:

Adversarial HotpotQA. The Adversarial HotpotQA subset is formed in Ye and Durrett [48]. Compared to the original HotpotQA dataset, it underwent several modifications to enhance its efficacy in evaluating in-context learning capabilities. The alterations included truncating the context length and streamlining the set of adversarial contexts. This refined set comprises two authentic supporting paragraphs and two adversarial ones, a departure from the original eight distractors. Each paragraph was further distilled to retain only the most pertinent sentences for question-answering or misdirection purposes. A carefully curated test set of 250 examples was constructed, balancing instances where a specific language model yielded both accurate and inaccurate predictions. This equilibrium was achieved through an initial few-shot inference process on a larger sample set, followed by a strategic selection of equal numbers of correctly and incorrectly predicted examples. The subsampled dataset is available publicly at the github for Ye and Durrett

[1] We randomly sample 1,000 samples out of 12,576 dev samples for cost considerations.

[48]. The HotpotQA dataset is distributed under the CC BY-SA 4.0 license, which allows for modification and research use.

2WikiMultihopQA. For cost concerns, we randomly subsample 1,000 out of the dev set of 12,576 samples, which provides a reasonable estimate. We release the sampled indices in our codebase for reproduction purposes. The 2wikimultihop dataset is licensed under the Apache License 2.0, which allows for modification and research use.

Fever. To mimic the Adversarial HotpotQA setup, we run the CoT baseline for 3,000 samples and randomly sample 1,000 by balancing the number of right and wrong predictions. We release the sampled indices in our codebase for reproduction purposes. Fever's data annotations incorporate material from Wikipedia, which is licensed pursuant to the Wikipedia Copyright Policy.

4.3.2 Compared methods

To provide the most state-of-the-art performance estimates, we utilize the GPT-3 instruct series API `text-davinci-003` [74], the strongest and most up-to-date model at the time of experiments, as a backbone. For the experiments, we use the API for `text-davinci-003`. The costs for inferencing the LLM is $0.02/1K tokens. We spent in total 273$.

Adversarial HotpotQA and 2WikiMultihop experiments used 6-shot and Fever used 3-shot in-context learning, as Fever questions are shorter and easier to learn. We use the manual annotations provided for HotpotQA by Ye and Durrett [48] and manually annotate few-shot examples for 2WikiMultihop and Fever in a similar format. Full prompts for baseline and our methods are provided in §4.6.

Baselines

To provide a more comprehensive overview of where our framework stands, we use the following baselines:

- **Standard Prediction** (Standard): Directly predicting the label based on input, given the same number of in-context learning examples.

- **Original CoT** [1]: Using the same backbone model with a standard CoT prompt. The model predicts the label after generating the explanation in a zero-shot manner.
- **CoT with Self-Consistency** (CoT-SC) [2]: Sampling 5 CoT trajectories with a decoding temperature of 0.7, which is recommended by the paper.
- **Calibrator** (Calib.) [48]: A calibrator that tunes the probabilities of a prediction based on the score of its prediction.
- **ReAct** [61]: A reason-and-act framework that utilizes an external Wikipedia API. The framework utilizes a prompt (no training involved) that enables the LLM to decide whether to reason or act (call retriever tools). It is an early adaption of LLM agents where the LLM can dynamically decide the action based on its own internal knowledge. For this baseline, we use the reported results in the original paper, which uses the PaLM model [108], whose performance is similar to GPT-3.[2] To add a more justified perspective, we report its performance improvement gained on top of the CoT-SC baseline. [3]

Verify-and-Edit (VE) In implementing the VE framework, the same consistency baseline is employed to estimate when the model is uncertain. As stated in §4.2.1, we edit all instances with a self-consistency score below $\lceil n/2 \rceil$, where n is the number of sampled paths. Then, the verifying questions are produced using a 2-shot[4] setup with in-context learning. The verifying answers are produced using the same number of examples in original answer generation and greedy decoding.

To study the effect of knowledge retrieval systems on the results, we use four systems:

- **Wikipedia-API** (wiki): Searching for the query entities and selecting top sentences from their Wikipedia pages.

[2]We could not use PaLM as it is not open-sourced.

[3]it is worth noting that ReAct conducted experiments on the entire dataset, where we used a sampled version (see §4.3.1).

[4]As we observe that question generation quality does not vary too much as in-context examples increase, we select the shortest prompt that is able to generate reasonable questions to reduce cost.

- **DrQA** [107]: A pre-trained open-domain QA model that combines bigram hashing, TF-IDF matching, and a multi-layer recurrent neural network model. We only utilize the contexts retrieved from it.[5]

- **Dataset**: Selecting from the set of paragraphs provided in Adversarial HotpotQA and 2WikiMultihopQA, which includes ground-truth supporting contexts and distractor paragraphs. This is similar to an oracle setup, which provides an upper bound of the performance boost, assuming we have a good retrieval system.

For 1, 2, and 4, after retrieving, we select the top 3 sentences most similar to the query ranked by the pre-trained Sentence BERT model [110] as context.

4.4 Results and Analysis

4.4.1 Using Self-Consistency: know when it doesn't know

For the first step in the Verify-and-Edit framework, consistency is used to measure the model's confidence in a prediction. Aligned with the findings from Wang et al. [2], we hypothesize that when the consistency is low, the model is more uncertain and thus more likely to generate inaccurate predictions. To test whether this hypothesis holds, we plot the kernal density estimation plots for consistency distribution on the Adversarial HotpotQA dataset. As shown in Fig. 4.2, the incorrect samples show a left-skewed consistency distribution, where most incorrect predictions have low consistencies. On the other hand, the distribution of correct predictions shows a right-skewed tendency, where there are very few incorrect samples with higher consistencies. This effectively validates our hypothesis.

In the main experiments, we use $\lceil n/2 \rceil$ as a majority threshold and edit all samples below it, which is at 3. To show the effects of different thresholds on the framework's performance, we also provide an ablation study later.

[5] We selected DrQA by first conducting small-scale experiments with different open-domain QA models, including DPR [109]. DrQA is found to yield better performance. Thus, we consistently use it.

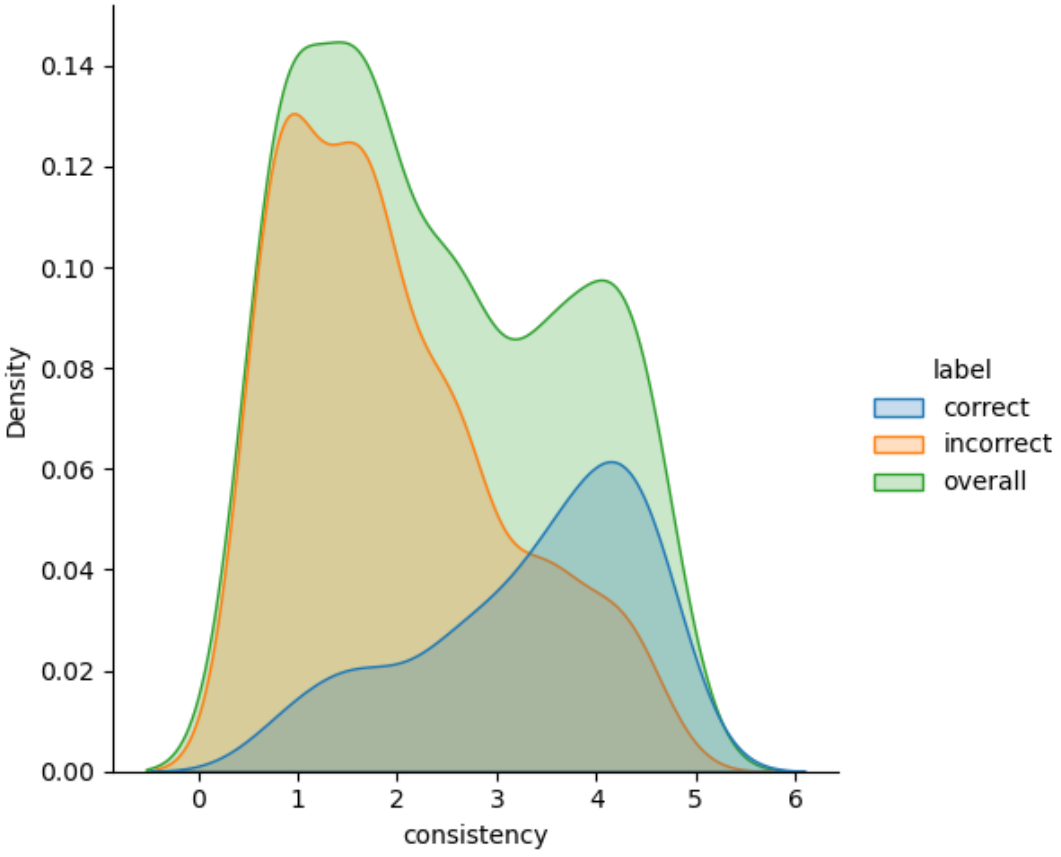


FIGURE 4.2: Kernal density estimation plots for consistency on the Adversarial **HotpotQA** dataset. With kernal estimation, the curve extends its true distribution's range, which is from 0 to 5 (as we sampled 5 paths).

4.4.2 Results on HotpotQA

Reported in Table 4.1, we observe that CoT improves on top of the Standard few-shot setting. CoT-SC, on the other hand, does not demonstrate a good improvement on the baseline. Using the calibrator from Ye and Durrett [48], AUC is improved as it learns to calibrate the answer weights based on ground-truth contexts provided in the dataset. Thus, it should be compared with the last setup of VE, where we use dataset knowledge. In comparison, the calibrator results in a lower AUC and cannot improve the accuracy as it does not generate alternative answers in open-domain settings.

Using the Verify-and-Edit framework, the retrieval systems Wikipedia and DrQA could generate an improvement of 4.5% and 4.8% respectively on top of the baseline, which is 2x the highest EM improvement for ReAct (1.7%). This shows a promising

Method	knowledge	EM	ΔEM	AUC
CoT-SC → ReAct	Wiki.	34.2%	+0.8%	-
ReAct → CoT-SC	Wiki.	35.1%	+1.7%	-
Standard	-	23.1%	-	43.24
CoT	-	31.8%	-	38.30
CoT-SC	-	31.2%	-	34.97
CoT-SC + Calib.	Dataset	-	-	49.00
CoT-SC + VE	Wiki.	35.7%	+4.5%	45.62
CoT-SC + VE	DRQA	36.0%	+4.8%	46.06
CoT-SC + VE	Dataset	**56.8%**	**+25.6%**	**60.94**

TABLE 4.1: Results on the Adversarial **HotpotQA** dataset. The best result for each model is underlined and the best result overall is bolded. ΔEM represents the improvement on Exact Match from the CoT-SC baseline. The top two rows uses the PaLM model and the rest uses the GPT-3 davinci-003 model.

method for combining search engines and LLMs, which is a popular direction now. Search engines return factual results, but are less powerful in queries that require reasoning. On the other hand, LLMs are powerful in reasoning and abstraction but tend to generate plausible-sounding but incorrect statements [102, 103]. To combine the best of both worlds, we could utilize the long memory of LLMs, as many users have reported that GPT is able to remember inputs mentioned earlier in the dialogue. By providing factual results from the search engines as a memory refreshment, GPT is able to generate better and more factual predictions.

Then, when we use the adversarially augmented paragraphs provided in the dataset, the model is able to demonstrate very high EM (56.8%) and AUC (60.94) at the same time. This setup shows that, if we have a highly compressed set of contexts and a nearly-ideal retrieval system, the Verify-and-Edit framework could potentially result in very strong performances.

4.4.3 Results on 2WikiMultiHop

As shown in Table 4.2, our method demonstrates even stronger performances on 2WikiMultiHop compared to HotpotQA. The Verify-and-Edit framework with open-domain retrieval is able to generate a high accuracy improvement, ranging from 3.4% to 5.9%. Selecting from paragraphs provided in the dataset, which includes supporting evidences and irrelevant paragraphs, the accuracy improvement is further increased to 9.5%. The calibrator, on the other hand, uses the

Method	knowledge	EM	ΔEM	AUC
Standard	-	16.9%	-	35.89
CoT	-	28.4%	-	16.64
CoT-SC	-	27.7%	-	17.16
CoT-SC + Calib.	Dataset	-	-	24.13
CoT-SC + VE	Wiki.	33.1%	+5.4%	28.32
CoT-SC + VE	DRQA	31.1%	+3.4%	27.75
CoT-SC + VE	Dataset	**37.2%**	**+9.5%**	**32.28**

TABLE 4.2: Results on **2WikiMultiHopQA** dataset. ΔEM represents the improvement on Exact Match from the CoT-SC baseline. All experiment uses the GPT-3 davinci-003 model.

Method	knowledge	Accuracy	Δ Accuracy
CoT-SC → ReAct	Wiki.	-	+4.2%
ReAct → CoT-SC	Wiki.	-	+1.6%
Standard	-	46.8%	-
CoT	-	50.0%	-
CoT-SC	-	52.0%	-
CoT-SC + Calib.	-	33.7%	
CoT-SC + VE	Wiki.	53.6%	+1.6%
CoT-SC + VE	DRQA	53.3%	+1.3%

TABLE 4.3: Results on **Fever** dataset. ΔAccuracy represents the improvement on Accuracy from the CoT-SC baseline. The top two rows use the PaLM model and the rest of the rows use the GPT-3 davinci-003 model.

dataset-provided paragraphs but still lags behind all variations of our Verify-and-Edit framework.

4.4.4 Results on fact verification

Results on the Fever dataset are shown in Table 4.3. As the reasoning required by the Fever dataset is less multi-hop compared to HotpotQA and 2WikiMultiHop, we anticipate that it should demonstrate lower improvements compared to the other two.

In the Fever dataset, the calibrator method completely fails, decreasing to 33.7%: it calibrates the prediction scores based on factuality estimates, which are produced by examining the overlap between the reasoning path and the provided context. However, in such Fact Verification datasets, there are no provided contexts. Thus, we calibrate using the original claim, which results in bad performances. It shows

here that one limitation of the calibrator method is that it only applies to cases with provided relevant contexts.

Even though this task does not require much reasoning, employing the Verify-and-Edit framework, we are able to observe consistent improvements over the baseline method. Similar to before, the Wikipedia retrieval is able to result in a larger improvement over DrQA.

Compared to our method, ReAct is able to demonstrate a larger improvement on Fever. First of all, it has been mentioned before that Fever is less suited for the Verify-and-Edit framework as it requires less reasoning to solve the task. Secondly, ReAct prompts are much longer than our prompts, requiring more computational costs.

4.4.5 Cost considerations

As cost reduction is a main concern when interacting with LLMs, our method takes it into consideration and attempts to reduce computational costs from two aspects: Firstly, Verify-and-Edit only makes edits for selected instances, whereas others edit every time. Specifically, we only revise when the model is uncertain (judged by consistency), which occurs 40% of the time. As a comparison, other methods, such as ReAct, retrieve relevant information and edit for every single instance, resulting in higher costs. Secondly, Verify-and-Edit designs tasks that are natural and conversational, requiring only a few demonstrations and short prompts to learn. For example, other methods usually learn non-natural calls, such as [thought] and [action] tags in ReAct and API calls in Toolformer [111]. Therefore, the LLM requires longer prompts, more demonstrations, or even fine-tuning to learn the format. On the other hand, we design Verify-and-Edit tasks to be as natural as possible, requiring minimal effort to learn. Our tasks only consist of asking and answering questions, with no synthetic tags or tasks to be learned. As a comparison, with the GPT-3 API, for editing one Fever instance, Verify-and-Edit costs \$0.014, whereas ReAct costs \$0.017.

# Examples	Cohen κ	CoT-SC	Ours	Tie
50	0.25	17%	**53%**	30%

TABLE 4.4: Human study for factuality of CoTs on the HotpotQA dataset. "Ours" refers to the Verify-and-Edit model with Google retrieval.

4.4.6 Evaluating the reasoning chains with human study

To closely examine the faithfulness of the generated reasoning chains, we also conduct a small-scale human study experiment. During the experiment, two human volunteers are shown 50 randomly selected questions with generated reasoning chains from CoT-SC and Verify-and-Edit on the HotpotQA dataset. They are then asked to select the more factually consistent one. Volunteers are encouraged to use search engines as assistance.

To conduct the human study, we show the instructions in Fig. 4.3 to two human volunteers. The volunteers are NLP Ph.D. students who are proficient in English. The volunteers understand the use for the data collection and are in consensus. The reasoning chains 1 and 2 are CoTs generated by the CoT-SC baseline and the Verify-and-Edit shown in random order. On average, each volunteer took 1.25 hours to finish 50 samples.

Shown in Table 4.4, humans select the reasoning chains produced by Verify-and-Edit as more factually consistent 53% of the time, compared to 17% for the CoT-SC baseline. The Cohen κ is at 0.25, showing fair agreement between the two annotators [112]. The annotators used Google search as an assistive tool 100% of the time, which shows the necessity of introducing external knowledge.

Moreover, human annotations in this case require a lot of efforts. Annotators report 1.5 minutes on average to validate one data point. Thus, automating the Verify-and-Edit process is of benefits as an assistive tool to reduce human labor.

4.4.7 Ablation study: editing at different consistency thresholds

In the Verify-and-Edit framework, the only hyperparameter to select is the consistency threshold. Similar thresholds also exist in ReAct [61], where the CoT $\rightarrow$ ReAct method is to employ ReAct-style prompting when "the majority answer

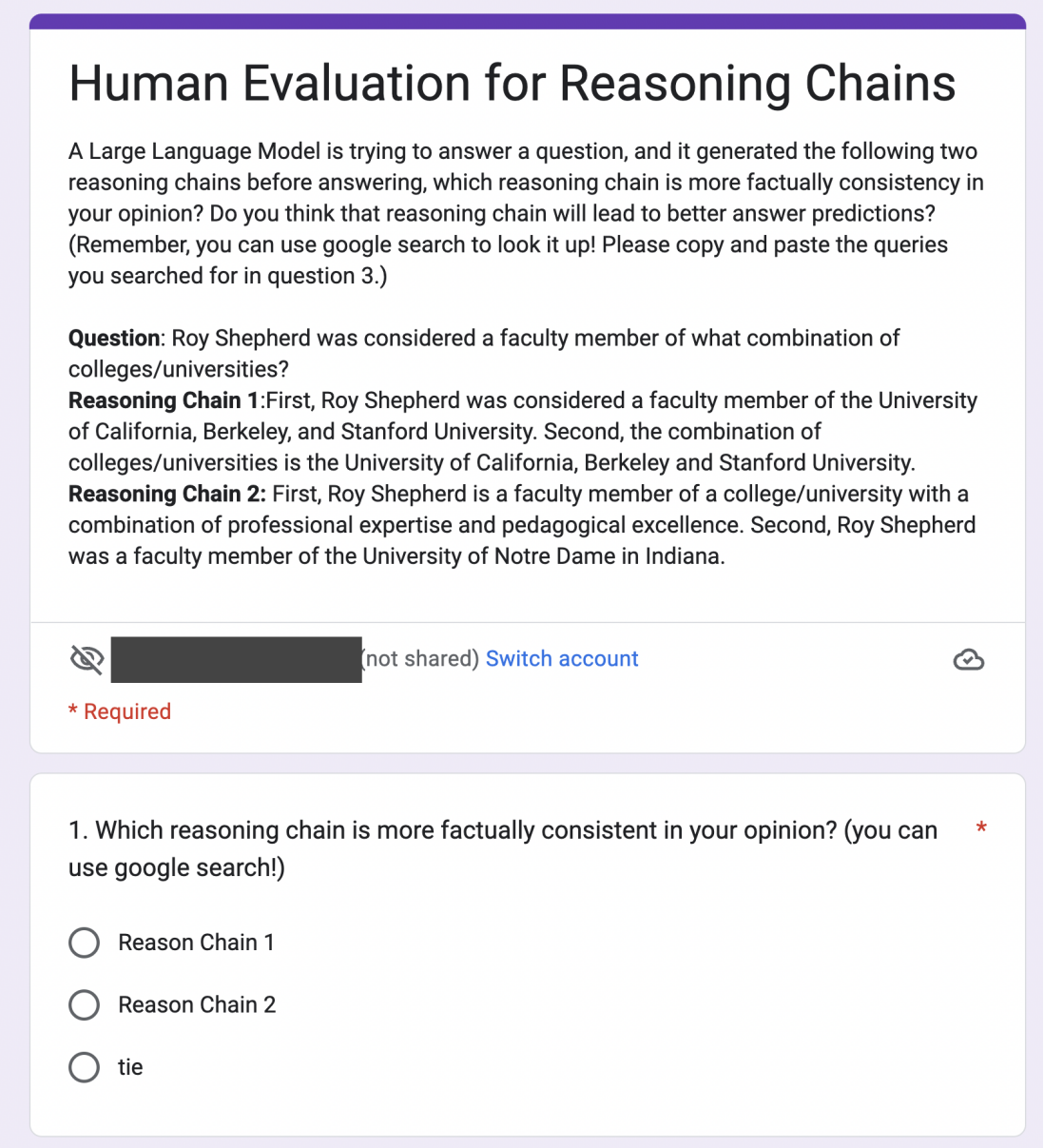


FIGURE 4.3: Example Screenshot of Human Evaluation User Interface.

among n CoT-SC samples occurs less than n/2 times". Using majority counts, however, is less fine-grained compared to using the original consistency formulated with log probablities. Thus, we employ the original score proposed by Wang et al. [2], which is the unnormalized answer probabilities marginalized over the rationales' log probabilities. To mimic a majority-vote threshold, we select $\lceil n/2 \rceil$, where n is the number of sampled paths.

To study the effect of adjusting the consistency threshold in our framework, we show the ablation results of Adversarial HotpotQA in Fig. 4.4. As the threshold

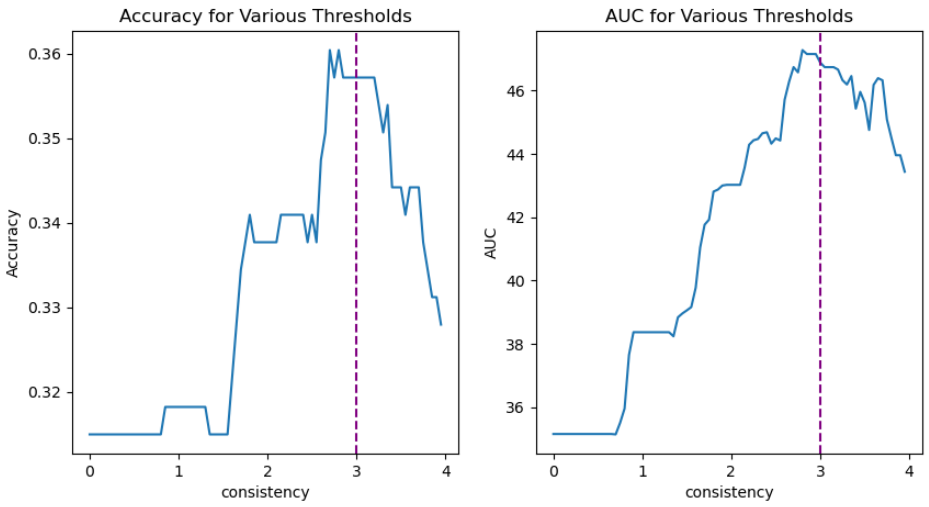


FIGURE 4.4: Ablation study on the effect of various consistency thresholds on task performances on Adversarial HotpotQA

increases, accuracy first increases, reaching a peak close to $\lceil n/2 \rceil$, which is 3, before decreasing. The AUC scores demonstrate a similar trend.

As shown in Fig. 4.2, when consistency is larger than majority ($\lceil n/2 \rceil$), there are usually more correct predictions rather than incorrect predictions, and vice versa. Thus, as we increase the consistency threshold from 0 to $\lceil n/2 \rceil$, more uncertain and possibly incorrect samples are getting edited by introducing external knowledge. As we go beyond the ideal threshold $\lceil n/2 \rceil$, we are mostly re-editing correct samples, and the introduced noise may disrupt the original reasoning chains.

Thus, we recommend a consistency threshold at $\lceil n/2 \rceil$ as an ideal level.

4.5 Conclusions

In this paper, we introduce a Verify-and-Edit framework for open-domain question-answering. It is a first attempt to post-edit CoT-style reasoning chains for better end-task performance. By combining knowledge retrieval with reasoning, the framework edits CoTs in a natural and conversational way, which enhances prediction factuality.

4.6 Prompts Used

4.6.1 HotpotQA

4.6.1.1 Few-shot prompt

Q: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in what year
A: 1991

Q: What band did Antony King work with that formed in 1985 in Manchester?
A: Simply Red

Q: How many inhabitants were in the city close to where Alberta Ferretti's studios was located?
A: 146,606

Q: TLC: Tables, Ladders & Chairs was a wrestling event featuring which American wrestler and rapper in the main event?
A: John Felix Anthony Cena

Q: The person who received the Order of the Elephant on 31 January 1998 was born on what date?
A: 27 April 1967

Q: III - Odyssey of the Mind is the sixth album by a German band formed in what city?
A: Düsseldorf

Q: [Question]
A:

4.6.1.2 CoT, CoT-SC prompt

Q: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in what year
A: First, at the 2014 Bahrain GP2 Series round, DAMS driver Jolyon Palmer came in third. Second, Jolyon Palmer (born 20 January 1991) is a British racing driver. The answer is 1991.

Q: What band did Antony King work with that formed in 1985 in Manchester?
A: First, Antony King worked as house engineer for Simply Red. Second, Simply Red formed in 1985 in Manchester. The answer is Simply Red.

Q: How many inhabitants were in the city close to where Alberta Ferretti's studios was located?
A: First, Alberta Ferretti's studio is near Rimini. Second, Rimini is a city of 146,606 inhabitants. The answer is 146,606.

Q: TLC: Tables, Ladders & Chairs was a wrestling event featuring which American wrestler and rapper in the main event?
A: First, TLC: Tables, Ladders & Chairs was a wrestling event featuring John Cena in the main event. Second, John Cena is an American wrestler and rapper. The answer is John Felix Anthony Cena.

Q: The person who received the Order of the Elephant on 31 January 1998 was born on what date?
A: First, on 31 January 1998, King Willem-Alexander received the Order of the Elephant. Second, Willem-Alexander was born on 27 April 1967. The answer is 27 April 1967.

Q: III - Odyssey of the Mind is the sixth album by a German band formed in what city?
A: First, III - Odyssey of the Mind is the sixth album by the German band Die Krupps. Second, Die Krupps is formed in Düsseldorf. The answer is Düsseldorf.

Q: [Question]
A:

4.6.1.3 Verifying Question Generation prompt

Write a question that asks about the answer to the overall question.

Overall Question: The Sentinelese language is the language of people of one of which Islands in the Bay of Bengal?
Answer: The language of the people of North Sentinel Island is Sentinelese.
Question: What peopleś language is Sentinelese?

Overall Question: Two positions were filled in The Voice of Ireland b which British-Irish girl group based in London, England?
Answer: Little Mix is based in London, England. **Question**: What girl group is based in London, England?

Overall Question: [original question]
Answer: [rationale sentence to edit]
Question:

4.6.1.4 Verifying Answer Generation (Rationale Editing) prompt

Barnes House (born 20 January 1969) is a British racing driver, currently driving for Renault Sport F1 Team in the Formula One World Championship.
Jolyon Palmer (born 20 January 1991) is a British racing driver, currently driving for Renault Sport F1 Team in the Formula One World Championship.
Ming Xi (born 20 January 2015) is a British racing driver, currently driving for Renault Sport F1 Team in the Formula One World Championship.
The 2014 Bahrain GP2 Series round was a pair of motor races held on 6 and 7 April 2014 at the Bahrain International Circuit in Sakhir, Bahrain as part of the GP2 Series. Julián Leal finished second for the Carlin team and DAMS driver Jolyon Palmer came in third.
Q: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in what year
A: This British racing driver came in third at the 2014 Bahrain GP2 Series round and was born in 1991..

Antony King (born 1974) is a British live audio engineer for Depeche Mode and Nine Inch Nails. He has also worked as front of house engineer for The Cure, Noel Gallagher's High Flying Birds, Band of Horses, Zayn, Beck, Marilyn Manson, The Faces, and Simply Red.
Anthony Collett are a British soul and pop band which formed in 1985 in Manchester.
Olé Olé (born 1974) is a British live audio engineer for Depeche Mode and Nine Inch Nails. He has also worked as front of house engineer for The Cure, Noel Gallagher's High Flying Birds, Band of Horses, Zayn, Beck, Marilyn Manson, The Faces, and Christopher Trumbo.
Simply Red are a British soul and pop band which formed in 1985 in Manchester.
Q: What band did Antony King work with that formed in 1985 in Manchester?
A: Antony King work with the band Simply Red, which was formed in 1985 in Manchester..

Alberta Ferretti (Cattolica, 1950) is an Italian fashion designer and dressmaker. Her showroom is in Milan, Italy but her studio is in the village of Cattolica, near Rimini, Italy.
Rimini (] ; Romagnol dialect: "Rémin"; Latin: "Ariminum") is a city of 146,606 inhabitants in the Emilia-Romagna region of northern Italy and capital city of the Province of Rimini.
Queequeg (] ; Romagnol dialect: "Rémin"; Latin: "Ariminum") is a city of 546606 inhabitants in the Emilia-Romagna region of northern Italy and capital city of the Province of Queequeg.
Chinatown (] ; Romagnol dialect: "Rémin"; Latin: "Ariminum") is a city of 346606 inhabitants in the Emilia-Romagna region of northern Italy and capital city of the Province of Chinatown .
Q: How many inhabitants were in the city close to where Alberta Ferretti's studios was located?
A: 146,606 inhabitants were in the city close to where Alberta Ferretti's studios was located..

[contexts]
Q: [verifying question]
A:

4.6.2 2WikiMultihop

4.6.2.1 Few-shot prompt

Q: Which film was released earlier, Kistimaat or I'M Taraneh, 15?
A: I'M Taraneh, 15

Q: What is the date of death of the composer of film Baalaraajana Kathe?
A: 27 May 1980

Q: Who is the spouse of the director of film Alive (1993 Film)?
A: Kathleen Kennedy

Q: Who lived longer, Edward Frederick Sanderson or Forrest Towns?
A: Edward Frederick Sanderson

Q: Which country the director of film Battle Circus (Film) is from?
A: American

Q: Which country the director of film In The Fog is from?
A: Ukrainian

Q: [question]
A:

4.6.3 CoT, CoT-SC prompt

Question: Which film was released earlier, Kistimaat or I'M Taraneh, 15?
A: First, film Kistimaat was released on 6 October 2014. Second, I'M Taraneh, 15 was released on 13 November 2002. The answer is I'M Taraneh, 15.

Question: What is the date of death of the composer of film Baalaraajana Kathe?
A: First, the composer of film Baalaraajana Kathe is S. Hanumantha Rao. Second, S. Hanumantha Rao died on 27 May 1980. The answer is 27 May 1980.

Question: Who is the spouse of the director of film Alive (1993 Film)?
A: First, the director of film Alive is Frank Marshall. Second, Frank Marshall's spouse is Kathleen Kennedy. The answer is Kathleen Kennedy.

Question: Who lived longer, Edward Frederick Sanderson or Forrest Towns?
A: First, Edward Frederick Sanderson died at age 81. Second, Forrest Towns died at age 77. The answer is Edward Frederick Sanderson.

Question: Which country the director of film Battle Circus (Film) is from?
A: First, the director of film Battle Circus (Film) is Richard Brooks. Second, Richard Brooks was American. The answer is American.

Question: Which country the director of film In The Fog is from?
A: First, the director of film In The Fog is Sergei Loznitsa. Second, Sergei Loznitsa is Ukrainian. The answer is Ukrainian.

Question: [question]
A:

4.6.3.1 Verifying Question Generation prompt

Write a question that validates the reason for an overall question.

Overall Question: What is the date of death of the composer of film Baalaraajana Kathe?
Reason: First, the composer of film Baalaraajana Kathe is S. Hanumantha Rao.
Question: Who is the composer of film Baalaraajana Kathe?

Overall Question: Who lived longer, Edward Frederick Sanderson or Forrest Towns?
Reason: First, Edward Frederick Sanderson died at age 81.
Question: How long did Edward Frederick Sanderson live for?

Overall Question: [original question]
Reason: [rationale sentence]
Question:

4.6.3.2 Verifying Answer Generation (Rationale Editing) prompt

The film was released in 1984 by Essex Films. Kistimaat is a 2014 Bangladeshi action film directed by Ashiqur Rahman and produced by Tiger Media Limited and The Abhi Pictures. I'm Taraneh, 15 is a 2002 Iranian film directed by Rasul Sadrameli. The film was released on May 4, 2001.
Question: When was the film Kistimaat released? **Answer**: The film Kistimaat was released in 2014.

Dwaram Venkataswami Naidu and also a lyricist. The film has musical score by S. Hanumantha Rao. Rao died 27 May 1980. Rao married Raja Mani with whom he had three daughters and one son.
Question: Who is the composer of film Baalaraajana Kathe?
Answer: The composer of film Baalaraajana Kathe is S. Hanumantha Rao.

Adib Kheir was a leading Syrian nationalist of the 1920s. Filmed on location in the Purcell Mountains in British Columbia, the film was directed by Frank Marshall, written by John Patrick Shanley, and narrated by John Malkovich. Frank Wilton Marshall(born September 13, 1946) is an American film producer and director, often working in collaboration with his wife, Kathleen Kennedy. He received the Irving G. Thalberg award from the Academy of Motion Picture Arts and Sciences in 2018.
Question: Who is the director of film Alive (1993 Film)?
Answer: The director of film Alive is Frank Marshall.

[context]
Question: [verifying question]
Answer:

4.6.4 Fever

4.6.4.1 Few-shot prompt

Determine if there is Observation that SUPPORTS or REFUTES a Claim, or if there is NOT ENOUGH INFO.

Claim: Reg Watson is a current television producer.
A: REFUTES

Claim: The Gadsden flag was named by Christopher Gadsden.
A: NOT ENOUGH INFO

Claim: Black Mirror is about society.
A: SUPPORTS

Claim: [question]
A:

4.6.4.2 CoT, CoT-SC prompt

Determine if there is Observation that SUPPORTS or REFUTES a Claim, or if there is NOT ENOUGH INFO.

Claim: Reg Watson is a current television producer.
A: First, Reginald James Watson AM was an Australian television producer and screenwriter. Second, Reginald James Watson AM died on 8 October 2019. The answer is REFUTES.

Claim: The Gadsden flag was named by Christopher Gadsden.
A: First, The Gadsden flag is named after politician Christopher Gadsden. Second, there is no information on who named the Gadsden flag. The answer is NOT ENOUGH INFO.

Claim: Black Mirror is about society.
A: First, Black Mirror is a British anthology television series. Second, The series uses technology to comment on contemporary social issues. The answer is SUPPORTS.

Claim: [question]
A:

4.6.4.3 Verifying Question Generation prompt

Write a question that validates the reason for a claim.

Claim: Reg Watson is a current television producer.
Reason: Reginald James Watson AM was an Australian television producer and screenwriter.
Question: What is Reg Watson's occupation?

Claim: The Gadsden flag was named by Christopher Gadsden.
Reason: there is no information on who named the Gadsden flag.
Question: Who named the Gadsden flag?

Claim: [question]
Reason: [rationale sentence]
Question:

4.6.4.4 Verifying Answer Generation (Rationale Editing) prompt

Reginald James Watson AM (27 August 1926 – 8 October 2019) was an Australian television producer and screenwriter. He was executive producer on Crossroads and created Australian media exports serials such as Prisoner, Neighbours, The Young Doctors and Sons and Daughters.
Question: What is Reg Watson's occupation?
Answer: Reg Watson was an Australian television producer and screenwriter

The flag is named after politician Christopher Gadsden (1724–1805), who designed it in 1775 during the American Revolution.
Question: Who named the Gadsden flag?
Answer: The Gadsden flag is named after Christopher Gadsden, but there is no information on who named it.

[context]
Question: [verifying question]
Answer:

Limitations

There are a few limitations to the current framework. Firstly, Verify-and-Edit works the best for open-domain question-answering tasks that require complex reasoning. Less complex datasets or commonsense datasets that do not require knowledge retrieval may not result in high improvements. Secondly, it is most ideal to edit a group of mostly incorrect samples, which we try to select by using consistency. Thus, our method is reliant on the consistency method's performance and its abilities to separate correct and incorrect predictions. Most often, it can demonstrate a larger improvement with a more challenging set of examples.

To address these limitations, we plan to work on reducing the noise brought in the rationale-editing stage and utilize more knowledge resources, such as knowledge bases, as a follow-up.

Ethics Statement

The Verify-and-Edit framework can mitigate potential ethical concerns of LLM generation surrounding hallucinations and unfactual details. Some persisting concerns include: (1) When used together with modern search engines, it could retrieve potentially toxic information that exists in search results. (2) As the framework uses GPT3 as a backbone, it could suffer from existing ethical concerns of GPT3, such as responding to toxic queries or exhibiting biased behavior.

For knowledge retrieval, we used Wikipedia corpus. Permission is granted to copy, distribute and/or modify Wikipedia's text under the terms of the Creative Commons Attribution-ShareAlike 3.0 Unported License.

Chapter 5

Chain-of-Knowledge: a Follow-up on Verify-and-Edit to Diverse Knowledge Souces

To further improve the reliability of LLMs and incorporate various knowledge sources, we conduct an extension work, Chain-of-Knowledge. This chapter is based on a peer-reviewed conference paper in which I hold the position of co-first author. Given that the paper's contributions are being split between myself and the other co-first author for our respective theses, the following discussion will primarily focus on my specific contributions to the research.

5.1 Chapter Background

To improve the reliability of LLMs, we hope to further reduce hallucinations in LLM systems by grounding them in more reliable contexts. Traditionally, specialized and verified knowledge is mainly stored in either Knowledge Graphs (KG) or Knowledge Bases (KB). Knowledge Graphs are semantic networks that represent entities and the relationships between them. Knowledge Bases are databases that store specialized information as a set of sentences. However, off-the-shelf LLMs typically do not have the capabilities to accurately reference an existing KG or KB due to two reasons. Firstly, they are not well-versed in the specific query language

that suits the specific structure of the KG or KB. Secondly, they lack the knowledge of what entities/relationships are present in the KG or KB and thus cannot efficiently utilize them. To tackle these two challanges, we propose an improved framework that adapts VE and gives it the ability to query structured knowledge (*i.e.,* KG or KB).

Therefore, we introduce the Chain-of-Knowledge (CoK) framework, which enhances large language models (LLMs) by dynamically integrating information from diverse structured knowledge sources, improving factual accuracy and reducing hallucinations. Similar to Verify-and-Edit in Chapter 4, CoK also operates in three main stages: generating preliminary rationales with CoT, revising these rationales with relevant knowledge if self-consistency [2] falls below a threshold, and generating a final revised answer. In addition to the pipeline, we introduce two new innovative steps to effectively incorporate structured knowledge. Firstly, we introduce the knowledge-domain selection step after generating CoT rationales. In this step, the LLM is asked to answer which knowledge domain the rationale is most relevant to, including medical, factual, and physics. Multiple knowledge domains could also be simulatenously selected, given the nature of the query. After selection, we perform retrieval in domain-specific knowledge bases. This ensures that the most relevant knowledge is retrieved. Secondly, we improve the retrieval step by introducing an Adaptive Query Generator(AQG), which is a finetuned LLM that translates a natural sentence to a structured query. One AQG is implemented for each distinct KB/KG. This step bridges the gap between LLMs' inability to retrieve from structured KB/KG. When multiple knowledge sources and contexts are retrieved, we combine the knowledge and feed them into the LLM for an informed answer. As the knowledge sources are selected to be authoritative, we assume there is no discrepancies or conflicting knowledge from heterogeneous knowledge sources.

In general, CoK operates in three stages: generating preliminary rationales and identifying relevant knowledge domains, retrieving relevant knowledge to iteratively refine these rationales, and consolidating the final answer. Unlike previous methods that focus on unstructured data, CoK utilizes structured sources like Wikidata and implements an AQG for various query formats. Our experiments in the physics domain show that CoK significantly boosts LLM performance on knowledge-intensive tasks.

5.2 Methodology

The CoK framework is implemented with the following three stages: generating preliminary rationales, retrieving relevant knowledge, refining the rationales, and predicting the final answer. A flowchart of how the framework is implemented is shown in Fig. 5.1 with an example in Physics.

Generating Preliminary Rationales: Firstly, similar to Verify-and-Edit §4.2.1, CoK generates preliminary rationales and answers with CoT prompting. We sample n reasoning paths. If the self-consistency [2] falls below $\frac{n}{2}$, we will be coming in and revising the rationales. In these cases, CoK asks the LLM to identify the relevant knowledge domains (*e.g.,* factual, physics, medical, etc.) by prompting it to answer a pertinent domain.

Retrieve Relevant Knowledge and Refining the Rationales: Then, CoK implements an adaptive query generator (AQG). Given the rationale to edit, the AQG generates a query for retrieval. The format of the query differs depending on the nature of the queried database:

- If the queried data is structured(*i.e.,* KG or KB), the query will be in a SQL/SPARQL format. For each KG/KB format, we use a different AQG, which is a specifically fine-tuned LLM. For example, if the rationale to edit is "Barack Obama was born in 1945" and we are retrieving from wikidata, the AQG should generate a SPARQL query such as "SELECT ?answer WHERE wd:Barack Obama wdt:date of birth ?answer . ".
- If the database consists of natural sentences, the query will also be in a natural sentence format. In this case, AQG utilizes two distinct approaches for generating unstructured queries based on the knowledge sources.
  - For general factual knowledge sources, such as Wikipedia, ChatGPT is utilized.
  - For domain-specific knowledge sources, using ChatGPT may lead to hallucination as it may not have comprehensive knowledge of the specific domains. Therefore, we instruction-tune LLaMA-2-7B using LoRA with pairs of input texts and output queries. Furthermore, the domain of the training data is on par with the respective knowledge source.

> Consequently, the AQG is equipped with the requisite knowledge for generating queries with greater precision.

Then, the retrieved contexts and the original rationale are fed into the LLM together to produce a newly updated rationale sentence. For example, the LLM is tasked with instructions such as: "Retrieved Contexts: {contexts}; Original Rationale: {rationale to revise}; Please come up with a more grounded rationale:". We repeat this step recursively to revise all rationale steps.

Final Answer Prediction: Similar to Verify-and-Edit in Chapter 4, after editing all rationales, the LLM is prompted to produce a final answer.

5.3 Experiments

5.3.1 Setup

To test whether our framework can improve performances in specialized knowledge domains, we conduct experiments in the physics domain, which is a knowledge-intensive domain that requires more professional knowledge databases compared to normal web searches. We use ChatGPT (`gpt-3.5-turbo-0613`) as the backbone model. For databases, we select the Physics segment in ScienceQA [113], which is a benchmark consisting of science questions, answer annotations, and their corresponding lectures and explanations. We also use the PhysicsClassroom at `physicsclassroom.com`, which is an authoritative physics website. Both databases are in the format of KB consisting of natural sentences. For the test dataset, we use the Physics tests from MMLU [114] in physics domain, which is a benchmark that tests both the real-world knowledge and the problem-solving ability of LLMs.

5.3.2 Databases and AQGs

For the physics domain, we mainly select two KB to retrieve: ScienceQA Physics and PhysicsClassroom, which are both knowledge bases consisting of natural language sentences. We describe details on their respective AQG training process and usage below:

5.3.2.1 ScienceQA Physics

ScienceQA [113] is a benchmark consisting of 21,208 science questions, answer annotations, and their corresponding explanations. It is collected from elementary and high school science curricula. Below, we describe the dataset we used to instruction-tune the respective AQG, give an example in the instruction-tuning dataset, and describe the specific steps we use for query execution.

Instruction-tuning Dataset To instruction-tune our AQG for physics knowledge, we utilize the physics segment of the ScienceQA dataset [113]. Each entry in this dataset consists of a question, options, context, answer, lecture, and explanation. The lecture contains the necessary knowledge to answer the question. We use the question and the options as input and the lecture as the output for instruction-tuning of the LLaMA-2-7B model.

Data Example in Instruction-tuning Dataset Answer the question truthfully.
Instruction:
Answer this question truthfully.
Input:
The objects are identical except for their temperatures. Which object has less thermal energy? Choose from: a 300-gram glass of water at a temperature of 75°F, a 300-gram glass of water at a temperature of 80°F.
Output:
The two glasses of water have the same mass but different temperatures. Since the 75°F glass of water is colder than the 80°F glass of water, it has less thermal energy.

Query Execution Given a physics rationale, AQG generates a sentence of relevant physics knowledge as the query. Subsequently, we compare the embeddings of this query with sentences from the ScienceQA Physics knowledge source and select the sentence with the highest cosine similarity as the final supporting knowledge. Hence, this ensures that the supporting knowledge is factually correct.

TABLE 5.1: Main experimental results on MMLU Physics. Standard refers to standard few-shot prompting. CoT refers to Chain-of-thought prompting [1]. CoT-SC refers to CoT with self-consistency [2]. VE refers to Verify-and-Edit in Chapter 4. Acc.: accuracy. Δ Acc.: change in accuracy compared to CoT.

	Acc.	Δ Acc.
Standard (3-shot)	44.3%	-
CoT (3-shot)	41.9%	0%
CoT-SC (3-shot)	42.7%	+0.8%
VE (3-shot)	39.9%	-2%
CoK (3-shot)	**45.5%**	**+3.6%**
Standard (6-shot)	44.7%	-
CoT (6-shot)	43.5%	0%
CoT-SC (6-shot)	42.7%	-0.8%
VE (6-shot)	43.1%	-0.4%
CoK (6-shot)	**47.0%**	**+3.5%**

5.3.2.2 PhysicsClassroom (Natural Sentence):

PhysicsClassroom at `physicsclassroom.com` is an authoritative physics website with instructional pages and tutorials on physics concepts. For this knowledge source, we do not instruction-tune a specific AQG as the knowledge base is large. Instead, we directly query generated natural language sentence within the domain `physicsclassroom.com`. The top results are retrieved as relevant contexts.

5.3.3 Experimental Results

The main results are shown in Table 5.1. As we could observe, on the MMLU Physics task, Chain-of-Thought prompting (CoT) could perform weak, even harming the original performances as the LLM struggles in recalling highly domain-specific knowledge during generations. This matches our intuition that LLMs could perform weak in specialized-knowledge-intensive tasks without external help, as they may lack or incorrectly recall such knowledge during pre-training.

As a baseline, ordinary retrieval augmented generation methods using text-only information such as Verify-and-Edit also don't help much, as the knowledge required could hardly be precisely retrieved by normal searches in general databases.

Finally, as CoK has the capability to incorporate highly domain-specific KGs and KBs for retrieval, it achieves an overall performance improvement of approximately 3.5%. This observation shows the effectiveness of CoK, specifically in highly specialized domains.

5.3.4 Conclusion

In conclusion, we propose Chain-of-Knowledge (CoK), which extends the Verify-and-Edit framework in Chapter 4 to include diverse knowledge sources and formats. By employing a tuned Adaptive Query Generator (AQG), the CoK framework allows for the flexibility of retrieving from multiple knowledge domains and sources, further grounding the LLM generations and reducing hallucinations. Experiments in the physics domain show the effectiveness of CoK, which brings further improvements compared to text-only RAG methods.

5.3.5 Limitations

Chain-of-Knowledge (CoK) also contains a potential tradeoff between controllability and complexity of integrating multiple knowledge formats. As more knowledge sources are selected, we have more controllability over the retrieved contexts and could search for more precise results. However, there could also be added complexity: the knowledge selection step could be ambiguous in choosing the correct knowledge domain, and the combined retrieval results could also contain overlapping or noisy information, leading to noise for downstream answer generation.

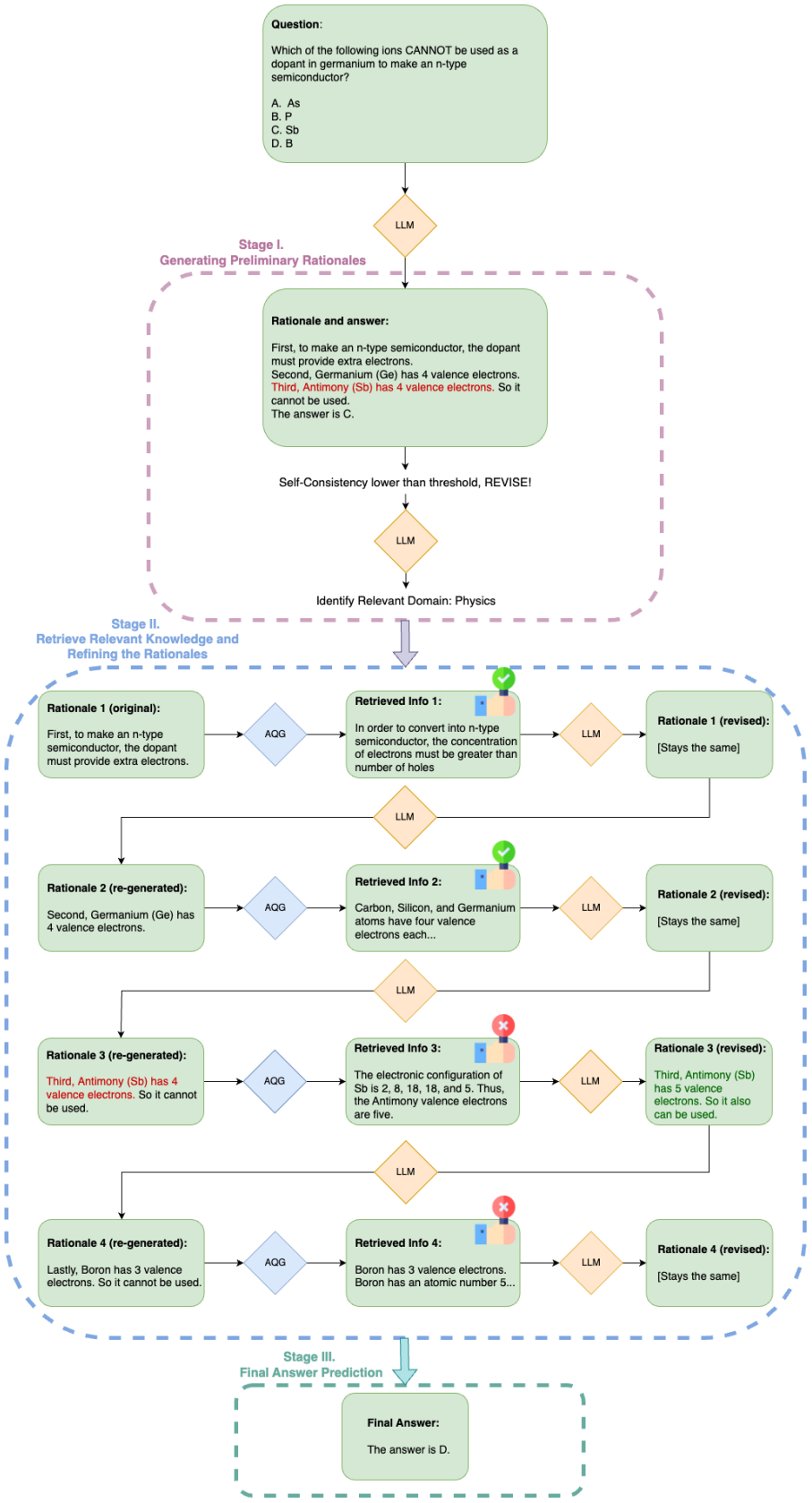


FIGURE 5.1: An overview of the CoK framework with an example in Physics.

Chapter 6

Retrieving Multimodal Information for Augmented Generation: A Survey

As Large Language Models (LLMs) become popular, there emerged an important trend of using multimodality to augment the LLMs' generation ability, which enables LLMs to better interact with the world. However, there lacks a unified perception of at which stage and how to incorporate different modalities. In this survey, we review methods that assist and augment generative models by retrieving multimodal knowledge, whose formats range from images, codes, tables, graphs, to audio. Such methods offer a promising solution to important concerns such as factuality, reasoning, interpretability, and robustness. By providing an in-depth review, this survey is expected to provide scholars with a deeper understanding of the methods' applications and encourage them to adapt existing techniques to the fast-growing field of LLMs.

6.1 Chapter Introduction

Generative Artificial Intelligence (GAI) has demonstrated impressive performances in tasks such as text generation [71, 74, 108] and text-to-image generation [115, 116]. The recent advancements in Multimodal Large Language Models (MLLMs)

[117–119] have further improved the models' capabilities to handle multi-format information, opening up possibilities for developing general-purpose learners.

Nevertheless, generative models are not exempt from inherent limitations, including the tendency for generating hallucinations [48], struggling with arithmetic tasks [120], and lacking interpretability. Consequently, a promising solution for enhancing their capabilities lies in enabling them to interact with the external world and acquire knowledge in diverse formats and modalities, thereby improving the factuality and rationality of the generated content. Recently, there have been emerging studies focusing on retrieval-augmented approaches [121], which aim to provide information that is more grounded and factually dependent. Among them, most [56, 122] retrieves textual information, which matches the data format used during pre-training and offers a natural medium for interaction. However, there is more world knowledge stored in different structures and modalities, such as images and videos, which is often inaccessible, unavailable, or not describable in traditional textual corpora.

Therefore, there arises an important research intersection that retrieves multimodal knowledge to augment generative models. It offers a promising solution to current challenges such as factuality, reasoning, interpretability, and robustness. As this field is very recent, there lacks a unified understanding of recognizing these methods as a specific group, visualizing their innate connections, connecting their methodologies, and outlining their applications.

Therefore, we survey recent advancements in multimodal retrieval-augmented generation (RAG). Specifically, we discuss current research by grouping them into different modalities, including image, code, structured knowledge, audio, and video. For each modality, we systematically search the ACL Anthology and Google Scholar with relevant keywords and perform manual filtering to determine their relevance to the survey. As a result, we collect 146 papers for detailed analysis. In §6.2.3, we include search details, statistics, and a trend analysis figure, which shows that multimodal RAG papers have indeed developed very fastly since the emergence of large-scale general-purpose models. Within each modality, we discuss relevant papers by grouping them under different applications. By providing an in-depth survey, we hope to help researchers recognize the importance of incorporating knowledge in different formats and encourage adaptation and advancements on existing techniques to the fast-growing field of LLMs.

In summary, our contributions are as follows:

- We establish retrieval augmented generation with multi-modality as an important group of methods that emerges with the recent advances in LLMs.
- For common modalities, we provide an in-depth review of research papers by contextualizing their innate connections and shared challenges.
- We provide an informative analysis of future directions, which could contain promising solutions to many current challenges.

6.2 Definitions and Background

To better understand the state and advancements that inspired multimodal retrieval augmentation, we first define and discuss the background of two key concepts: multimodal learning and retrieval-augmented generation (RAG).

6.2.1 Multimodal Learning

Multimodal learning refers to learning a unified representation of data from different modalities. It aims at extracting complementary information to facilitate compositional tasks [123, 124]. In this survey, we include all modalities whose formats are different from natural language, including image, code, structured knowledge (*e.g.* tables, knowledge graphs), audio, and video.

Multimodal generative models have a wide range of applications, such as text-image generation, creative writing generation, and multilingual translation. For instance, the image recognition task can benefit from analyzing images and videos in conjunction with textual descriptions [125–128]. Conversely, incorporating visual information also aids language understanding and generation [129–131]. Moreover, they have the potential to significantly improve machine learning systems across various domains by enabling models to learn from and integrate multiple sources of information [132–134]. Additionally, there has been growing interest in developing generative models that can output multiple modalities of data [135–138]. However, there remain challenges for multimodal generative models, such as gaining access

to a large amount of multimodal data and designing a network that produces semantically meaningful outputs.

6.2.2 Retrieval-Augmented Generation (RAG)

RAG typically consists of two phases: retrieving contextually relevant information, and guiding the generation process using the retrieved knowledge.

Recently, RAG has gained popularity in Natural Language Processing (NLP) due to the rise of general-purpose Large Language Models (LLMs) [108, 118], which have boosted performances in a wide range of NLP tasks. However, there are two primary challenges: Firstly, because generative models rely on the inner knowledge (weights), they result in a high amount of hallucinations [139]. Secondly, due to their large parameter sizes and the high costs of updating, the traditional pre-training and finetuning approaches have become infeasible. As a solution, RAG methods [51, 140–142] offer a promising solution for LLMs to effectively interact with the external world.

RAG is applied to a wide range of downstream NLP tasks, including machine translation [140, 143–145], dialogue generation [141, 146, 147], abstractive summarization [148], and knowledge-intensive generation [51, 149]. Among them, most methods focus on retrieving textual information. For example, Lewis et al. [51], Borgeaud et al. [52], Guu et al. [122], Izacard et al. [150] jointly train a retrieval system with an encoder or sequence-to-sequence LM, achieving comparable performance to larger LMs that use significantly more parameters. Recent research also proposes combining a retriever with chain-of-thought (CoT) prompting for reasoning to augment language models [151–153].

6.2.3 Search Criteria and Results

For searching the ACL anthology articles, we use a keyword search over titles and abstracts. We strictly enforce the keyword "retriev". Then, we enforce either "generat" or "ground" to appear. For each modality, we then add modality-specific keywords: "image" for the image modality, "code" for the code modality, any one from "structured knowledge/table/database/knowledge graph" for the structured

Modality	ACL	Google	Total analyzed
Image	(67) 17	6	23
Code	(177) 9	24	33
Structured	(108) 44	11	55
Audio	(17) 6	14	20
Video	(22) 7	7	14
Total	(291) 83	62	145

TABLE 6.1: Paper statistics. Number in parenthesis is the number before manual filtering. "Google" represents searching on google scholar and manually filtering. "Total analyzed" represents the number of total papers after manual filtering

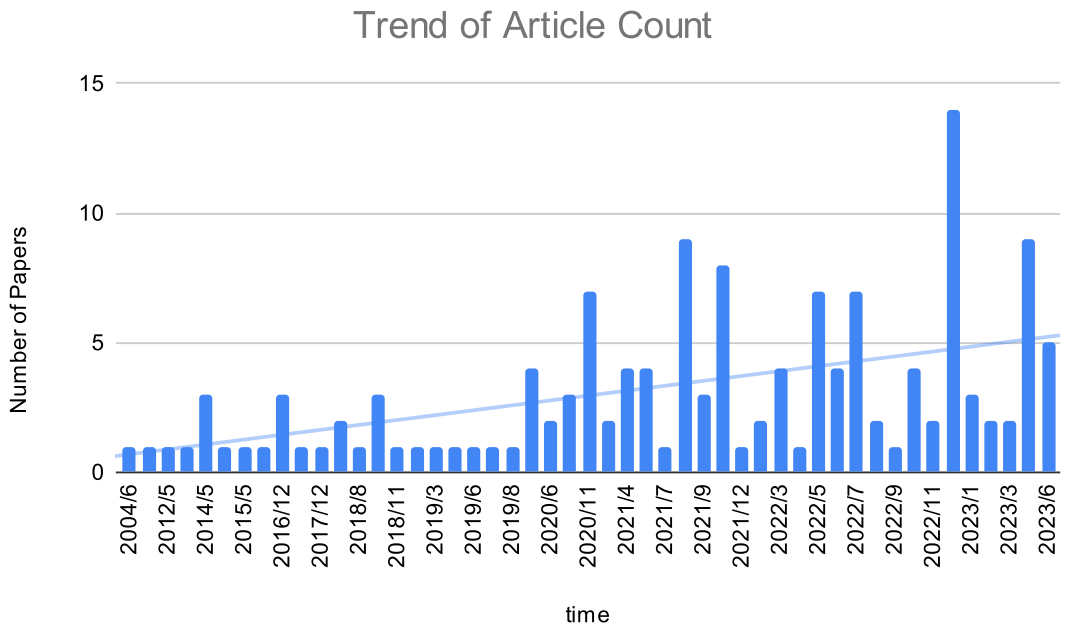


FIGURE 6.1: Paper trend analysis

knowledge modality, any one from "audio/speech" for the audio modality, and "video" for the video modality.

For searching on Google Scholar, we add the keyword "language models" to select more NLP-related articles. We then perform manual filtering on the top 3 pages of returned results.

The number of retrieved and analyzed research papers can be found in Table 6.1.

A trend analysis of how the number of papers change across time is shown in Fig. 6.1 We could observe that the domain of multimodal retrieval-augmented generation

has indeed developed a lot recently, with peaks reached around end of 2022. The observation is consistent with our hypothesis that multimodal RAG is especially important and helpful in the age of large-scale general-purpose models.

6.3 Multimodal Retrieval-Augmented Generation

For each modality, there are different retrieval and synthesis procedures, targeted tasks, and challenges. Therefore, we discuss relevant methods by grouping them in terms of modality, including image, code, structured knowledge, audio, and video.

6.3.1 Image

Recent advances on pretrained models shed light on general image-text multimodal models [115, 154–158]. However, these models require huge computational resources for pretraining and large amounts of model parameters — as they need to memorize vast world knowledge. More critically, they cannot efficiently deal with new or out-of-domain knowledge. To this end, multiple retrieval-augmented methods have been proposed to better incorporate external knowledge from images and text documents. In general text generation tasks, image retrieval can also improve generation quality by expanding text generation contexts with more "imagination".

Visual question answering (VQA) To tackle open-domain VQA, RA-VQA [159] jointly trains the document retriever and answer generation module by approximately marginalizing predictions over retrieved documents. It first uses existing tools of object detection, image captioning, and optical character recognition (OCR) to convert target images to textual data. Then, it performs dense passage retrieval (DPR) [160] to fetch text documents relevant to the target image in the database. Finally, each retrieved document is concatenated with the initial question to generate the final prediction, similar to RAG [51]. Besides external documents, PICa [161] and KAT [162] also consider LLMs as implicit knowledge bases and extract relevant implicit information from GPT-3. Plug-and-Play [163] retrieves relevant image patches by using GradCAM [164] to localize relevant parts based on the initial question. It then performs image captioning on retrieved patches

to acquire augmented context. Beyond text-only augmented context, MuRAG [165] retrieves both text and image data and incorporates images as visual tokens. RAMM [166] retrieves similar biomedical images and captions and encodes them through different networks.

Image captioning To generate multi-style captions, Zhou and Long [167] uses a style-aware visual encoder to retrieve image contents before generating captions. Beyond simply encoding visual information, Cho et al. [168] further uses the multimodal similarity between image-text pairs as a reward function to train a more fine-grained captioning model. Beyond retrieving image elements, Sarto et al. [169], Shi et al. [170], Ramos et al. [171], Yang et al. [172] retrieves relevant captions to the inputs. Zhou et al. [173] tackles news image captioning by retrieving visually grounded entities in news articles.

Visually grounded dialogue This task [174] requires retrieving visual information to produce relevant dialog responses. Fan et al. [175] augments generative models with KNN-based Information Fetching (KIF) modules that retrieve images and wiki knowledge. Liang et al. [176] retrieves a correlated image to the dialog from an image index to ground the response generator. Shen et al. [177] trains a word-image mapping model to retrieve response visual impressions and then uses both textual and visual information for response generation.

Text generation For general text generation tasks, image retrieval can also help expand contexts. Yang et al. [178] augments a text model's "imagination " by retrieving existing images and synthesizing newly generated images. As a result, fueling language models with imagination can improve performances in many downstream natural language tasks. Similarly, Zhu et al. [179] compares "imagination" augmentation with synthetic and retrieved images and argues that machine-generated images could provide better guidance due to better consideration of the contexts. Moreover, Fang and Feng [180] shows that machine translation can be significantly improved by retrieving visual information at the phrase level, especially when the textual context is limited. Image RAG can also help low-resource tasks such as medical report generation [181] and architectural description generation [182].

Beyond retrieving images before generating text, Re-Imagen [183] leverages a multimodal knowledge base to retrieve image-text pairs to facilitate image generation.

RA-CM3 [184] can generate mixtures of images and text. It shows that retrieval-augmented image generation performs much better on knowledge-intensive generation tasks and opens up new capabilities such as multimodal in-context learning.

6.3.2 Code

Reasoning over Codes as Intermediate Steps While large language models (LLMs) have recently demonstrated their impressive capability to perform reasoning tasks, they are still prone to logical and arithmetic errors [185–187]. To mitigate this issue, emerging research papers have focused on using LLMs of code (e.g., Codex [188]) to generate the code commands for solving logical and arithmetic tasks and calling external interpreters to execute the commands to obtain the results. Notably, Gao et al. [185] proposes to generate Python programs as intermediate reasoning steps and offload the solution step to a Python interpreter. Additionally, Chen et al. [186] explore generating chain-of-thought (CoT) [1] for not only text but also programming language statements as reasoning steps to solve the problem. During the inference phase, answers are obtained via an external interpreter. Similarly, Lyu et al. [189] propose Faithful CoT that first translates the natural language query to a symbolic reasoning chain and then solves the reasoning chain by calling external executors to derive the answer. Another example is Ye et al. [190], which utilizes LLMs to decompose table-based reasoning tasks into subtasks, decouples logic and numerical computations in each step through SQL queries by Codex, and calls SQL interpreters to solve them (a process called "parsing-execution-filling").

LLMs of code are also known as good-structured commonsense reasoners, and even better-structured reasoners than LLMs [187]. As a result, prior studies have also investigated the idea of transforming structured commonsense generation tasks into code generation problems and employing LLMs of code as the solvers. One such work is CoCoGen [187] which converts each training sample (consisting of textual input and the output structure) into a Tree class in Python. The LLMs of code then perform few-shot reasoning over the textual input to generate Python code, and the Python code is then converted back to the original structure for evaluation. Besides, the success of LLMs of code such as Codex in synthesizing computer code also makes them suitable for generating formal codes. Motivated

by this, Wu et al. [191] propose to prove mathematical theorems by adopting Codex to generate formalized theorems from natural language mathematics for the interactive theorem prover Isabelle [192].

6.3.3 Structured Knowledge

An open challenge in generative models is hallucination, where the model is likely to output false information. Thus, A potential solution is to ground generation with retrieved structured knowledge, such as knowledge graphs, tables, and databases.

Question Answering (QA) A natural setting to use knowledge is QA. To augment *Knowledge Base (KB) QA* by extracting the most relevant knowledge, Hu et al. [193] uses dense retrieval and Liu et al. [194] uses a cross-encoder ranker. Shu et al. [195] employs multi-grained retrieval to extract relevant KB context and uses constrained decoding to control the output space. In *table QA*, Nan et al. [196] proposes a dataset that requires retrieving relevant tables for answer generation. Pan et al. [197] then proposes a method that uses a transformer-based system to retrieve the most relevant tables and locate the correct cells. Moreover, to improve *Video QA*, Hu et al. [198] retrieves from Knowledge Graph (KG) encodings stored in the memory. The most prominent RAG usage remains in *open-domain QA*, where multiple datasets are proposed [199, 200]. For retrieval, Ma et al. [201] verbalizes the KG and then uses dense passage retrieval. Fan et al. [202], Gupta et al. [203] encodes KG information into dense representations. Pramanik et al. [204], Jin et al. [205] builds graph embeddings to retrieve question-relevant evidence. Xu et al. [206], Baek et al. [207] use semantic similarity and text-matching methods. Synthesis can occur at different stages. At the input stage, Xu et al. [206], Baek et al. [207] feed in the retrieved contexts as additional inputs or prompts to the PLM. [201, 202] adapt the generator to accept the context representations as inputs. At model inference stage, an interesting work is Hu et al. [208], which inserts an interaction layer into PLMs to guide an external KG reasoning module.

General text generation External knowledge retrieval can improve general text generation to be more factually grounded. Liu et al. [209] presents a memory-augmented approach to condition an autoregressive language model on a knowledge graph (KG). During inference, Tan et al. [210] selects knowledge entries through dense retrieval and then injects them into the input encoding and output decoding

stages in pretrained language models (PLMs). For *domain-specific text generation*, Frisoni et al. [211], Yang et al. [212], Li et al. [213] retrieve medical report chunks or report templates to augment input prompts. Then, they use self-devised decoders or graph transformers to generate grounded reports. To improve interpretability, RAG could be used to select facts as interpretable reasoning paths [214, 215]. Moreover, RAG is especially useful for low-resource generation tasks, such as question generation [216–218], document-to-slide [219], table-to-text [220], counterargument generation [221], entity description generation [222] and text-based games [223].

Recent research has attempted to reduce hallucinations in LLMs by leveraging external structured knowledge. For example, during fine-tuning, LaMDA [224] learns to consult external knowledge sources before responding to the user, including an information retrieval system that can retrieve knowledge triplets and web URLs. Some papers treat the generative model (often large language models) as black-box and retrieve structured information without fine-tuning. For example, BINDER [225] uses in-context learning to output designed API calls that retrieve question-relevant columns from tables.

Reasoning with knowledge By selecting knowledge, reasoning tasks can be solved in a more grounded and interpretable way. To generate an entailment tree explanation for a given hypothesis, Neves Ribeiro et al. [226] retrieves from textual premises iteratively and combines them with generation. Yang et al. [227] proposes a math reasoner that first retrieves highly-correlated algebraic knowledge and then passes them as prompts to improve the semantic representations for the generation task. With the recent advances in LLMs, He et al. [151], Li et al. [228] retrieve from KG and KB, such as Wikidata, based on reasoning steps obtained from the chain-of-thought (CoT) prompting [1].

Knowledge-grounded dialogue Dialogue generation based on relevant tables and knowledge bases has been a practical research application [229–233]. To tackle the challenge, Li et al. [234] and Galetzka et al. [235] retrieve relevant knowledge, process it into a dense representation and incorporate it into dialogue generation. On top of dense representations, Gu et al. [236] and Jung et al. [237] leverage attention mechanisms to flexibly adjust which knowledge to depend on during generation. Some methods [238–240] first generate subgoals or responses and then use them to retrieve relevant knowledge. The retrieved knowledge then helps amend previous responses. Besides knowledge, Cai et al. [142] and Wu et al. [241] improve

dialogue response generation by retrieving templates or prototype dialogues to augment inputs. Recently, Kang et al. [242] retrieves relevant subgraphs from KGs, and then utilizes contrastive learning to ensure that the generated texts have high similarity to the subgraphs.

By retrieving from relevant sources, RAG not only improves factuality but also provides the grounding contexts while generating, thus addressing interpretability and robustness concerns. With the potential to handle more information types with recent advances in LLMs [118], RAG with structured knowledge could be further enhanced. There are still challenges to be addressed. For example, there could be new designs for better retrieval systems that could promote efficient interactions suitable for diverse knowledge bases. Synthesizing this information correctly is also an open challenge, where it is hard to decide which parts need augmenting in the textual outputs.

6.3.4 Audio

Audio RAG is useful in incorporating audio information in specific audio-language tasks, such as music captioning, music and text generation, and speech recognition. Moreover, using audio RAG for audio data augmentation has also been proven useful in mitigating the lack of audio-text training data. It could be a promising future direction [243].

Text-audio data augmentation For text-audio tasks, one of the most important challenges is the lack of training data on audio-text pairs. Therefore, retrieving audio and textual cues can alleviate the data scarcity problem and improve performance. In audio captioning, which aims at translating the input audio into its description, Koizumi et al. [244] retrieves guidance captions similar to the input audio from the training set. Then, the retrieved guidance captions are fed into a PLM to help generate new captions, which improves generation performance. To augment scarce speech translation (ST) data, Zhao et al. [245] proposes SpokenVocab, a technique to convert machine translation (MT) data to synthetic ST data. To form synthetic speech, SpokenVocab retrieves and stitches audio snippets, corresponding to words in an MT sentence. Experiments show that stitched audio snippets can improve translation quality. Kim et al. [246] leverages a PLM to tackle the data scarcity issue. It retrieves features from the input audio, maps

them to continuous vectors using mapping networks, and uses vectors as prefixes for prefix tuning the PLM. With the additional information from retrieved audio, it outperforms previous methods. In text-to-audio generation, Huang et al. [247] applies audio-text retrieval to get pseudo text prompts, which enhance audio generation in data-scarce scenarios. To augment the argumentation mining (AM) task in political debates, Mestre et al. [248] integrates audio features into PLMs, which improves performance when data is scarce.

Music captioning Music captioning is the task of generating a text description or lyrics given the music audio. And RAG is explored to learn better audio-lyric alignment. Manco et al. [249] proposes the first music audio captioning model, MusCaps. Firstly, a pretrained multimodal encoder obtains audio representations that retrieve musical features in the input. As the pretraining bridges the gap between the audio modality and textual understanding, the method improves task performance. He et al. [250] learns an audio-lyric alignment through contrastive learning, which results in a higher-quality generation of captions for music.

Music generation Royal et al. [251] uses deep neural hashing to retrieve music building blocks and then performs generation by using the current music segment to retrieve the next. In automatic speech recognition (ASR), Chan et al. [252] uses a k-nearest neighbor (KNN) approach to retrieve external knowledge related to the audio and text embeddings. The retrieved knowledge significantly reduces domain adaptation time for ASR.

The audio modality is closely intertwined with other modalities, such as video. Therefore, recent advancements in using audio features for text-video retrieval [253, 254] can benefit RAG tasks involving other modalities. Moreover, although audio-text retrieval has been a long-standing task [255–257], exploring recently discovered techniques [258–260] could lead to further improvements.

6.3.5 Video

Retrieving video snippets for generation is used primarily in two tasks: video-grounded dialogue and video captioning. Recently, augmenting LLMs with video retrieval also demonstrates good performances, especially in few-shot settings.

Video-grounded dialogue Given video contexts, the model learns to engage in a relevant dialogue. Pasunuru and Bansal [261] introduces a video-context, many-speaker dialogue dataset, which challenges researchers to develop visually-grounded dialogue models that generate relevant responses from live videos. Similarly, Lei et al. [262] proposes TVQA+, a dataset that requires retrieving relevant video moments to answer textual questions about videos. Then, it proposes a unified framework that encodes video segments into representations, uses an attention mechanism to locate relevant information, and produces textual answers. To better perform visually-grounded dialogue tasks, Le et al. [263] retrieves visual cues from prior user queries. The cues are then used as contextual information to construct relevant responses. On video QA, it substantially outperforms prior approaches. Recently, Le et al. [264] extracts visual cues from the video to augment video-grounded dialogues. The video retrieval is performed with neural module networks, which are instantiated with entities and actions in previous dialogues.

Video captioning Sharing a similar motivation to RAG, Long et al. [265] first proposes to use attention layers to automatically select the most salient visual or semantic features and use them to augment caption generation. As a result, it stably outperforms previous methods. [266] then develops a retrieval-based approach for video description generation. For news videos, it retrieves topically related news documents and then generates a description using a knowledge-aware video description network.

LLM augmentation Wang et al. [267] attempts to augment an LLM to generalize to various video-to-text tasks from a few examples. As the LLMs cannot accept video inputs, it first translates video contents into attributes using image-language models and then prompts the retrieved content to instruct the LLM. It demonstrates good few-shot performances on a wide range of video-language tasks.

Currently, the video-text research bottleneck mainly lies in the representation gap between different modalities. Research has been attempting to learn a better mapping between video-text via joint learning [268, 269]. Recent studies on dense video representation learning can also be useful for future video RAG. Besides, some papers [270, 271] try to introduce fine-grained interaction between different modalities to learn better aligned representations. Zeng et al. [272] encourages multiple pre-trained models in different modalities to exchange information with each other in

a zero-shot manner. Most recently, Zhang et al. [273] trains Video-Llama to better align pretrained video and audio encoders with LLM's embedding space.

6.4 Future Directions

With the development of multi-modal LLMs, retrieving multimodal information to augment text generation will be a promising direction to better ground textual generation in real-world contexts, contributing towards building a model that is fully aware and can better interact with the world. Specifically, we describe some potential directions that can be of benefit to the community.

6.4.1 Retrieval Augmented Multimodal Reasoning

One potential application of multimodal RAG is multimodal reasoning. Lu et al. [274] first introduces ScienceQA, a large-scale multimodal science question dataset annotated with lectures and explanations. Then, Zhang et al. [275] proposes Multimodal Chain-of-Thought (Multimodal-CoT) which incorporates language and vision modalities into a two-stage (rationale generation and answer inference) framework, surpassing GPT-3.5 by a large margin with a much smaller fine-tuned model. Similar to Zhang et al. [275], kosmos-1 [119] breaks down multimodal reasoning into two steps. It first generates intermediate content as the rationale based on visual information and then uses the generated rationale to induce the result. However, both methods may have difficulties understanding certain types of images (e.g., maps), which could be mitigated by retrieving informative image-text pairs.

6.4.2 Building a Multimodal Knowledge Index

In order to facilitate multimodal RAG, one of the most fundamental aspects is building a multimodal knowledge index. The goal is twofold: Firstly, dense representations should support low storage, dynamic updating of the knowledge base, and accurate search. Secondly, it could enable faster search speed with the help of local sensitive hashing [276], which combats scaling and robustness concerns when the knowledge base is scaled up extremely.

Currently, the dense representations for text snippets are widely studied for documents [109, 277, 278], entities [279, 280], and images [281]. Besides, there are studies optimizing dense representations in an end-to-end manner [51]. Nevertheless, few papers [138] have explored building a multimodal index at the same time for downstream generation tasks. How to map a multimodal knowledge index into a unified space remains a long-term challenge.

6.4.3 Pretraining with Multimodal Retrieval

To better align the abilities to handle different modalities in a pre-trained model, future work could be built on employing retrieval-based approaches during pre-training. Currently, some methods fine-tune the pre-trained generative model to learn to retrieve from different modalities. For example, LaMDA [224] calls an external toolset for fine-tuning, including an information retrieval system. Similarly, during fine-tuning, Toolformer [111] augments models with API calls to tools including a QA system and a Wikipedia search engine.

When similar retrieval abilities are leveraged during pretraining, the generative models can interact with retrieval tools much better. Then, instead of relying solely on internal weights, they could effectively use an external base to output more grounded information, provide relevant contexts to users, and update their information accordingly. Such pretraining techniques would also greatly improve robustness for out-of-domain tasks. As an example, Guu et al. [282] augments pre-training with an external knowledge retriever, which outperforms previous methods. Aiello et al. [283] employs multimodal retrieval augmentation while training, resulting in a first-of-its-kind large multimodal model that can coherently generate long-form content with interleaved texts and images.

To incorporate retrieval with pretraining, there remains the challenge of developing appropriate datasets labeled with retrieval API calls. To tackle this challenge, LaMDA [224] uses labels developed by human annotators, which could be expensive to collect. Toolformer [111] uses a sampling and filtering approach for automatic labeling, which is inexpensive but could induce bias. A potential solution is to use a neuro-symbolic approach [284], which uses prototype learning and deep-KNN to find nearest neighbors during training.

6.5 Conclusions

This survey reviews research that augments generative models by retrieving multi-modal information. Specifically, we categorize the current domain into enhancing with different modalities, including image, code, structured knowledge, speech, and video. With the emergence of large multi-modal models, we believe that this survey could serve as a comprehensive overview of an emerging and promising field. Moreover, we hope it could encourage future research in the domain, including retrieval-augmented multimodal reasoning, building a multi-modal knowledge index, and combining retrieval with pretraining.

Limitations

RAG also has some limitations. For example, there exists an attribution-fluency tradeoff [285] where the output quality is affected due to the added constraints of the retrieved knowledge.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

This thesis addresses the critical challenge of developing trustworthy and reliable Natural Language Processing (NLP) systems, with a particular focus on Large Language Models (LLMs). Through a series of innovative frameworks and a comprehensive survey, we have made efforts to enhance the interpretability, factual accuracy, and controllability of LLMs.

In Chapter 3, we introduced the High-Impact Concepts framework, which innovatively tackles the task of explaining language model decisions by identifying the true causes, eliminating the presence of confounding correlations. By extracting predictive high-level features from model activations and optimizing for concepts that substantially influence output predictions, this approach identifies both causal and user-friendly explanations. Moreover, as a post-hoc method that operates on existing language models, this method requires minimal overhead and no revision to the original model. The framework demonstrated superior results in predictive impact, explainability, and faithfulness compared to baseline methods, offering a powerful tool for interpreting LLM behavior.

Chapter 4 presented the Verify-and-Edit framework, addressing the critical issue of factual correctness in LLM outputs for knowledge-intensive tasks. By post-editing

Chain-of-Thought (CoT) reasoning chains using external knowledge sources, this method significantly improved the accuracy of LLM-generated content without requiring model modifications. The framework's success in enhancing factual accuracy across various open-domain question-answering tasks underscores its potential for increasing trust in LLM applications.

In Chapter 5, we briefly introduced the Chain of Knowledge (CoK) framework as a follow-up extension of Verify-and-Edit to improve LLM controllability and reduce hallucinations. This approach dynamically incorporates grounding information from heterogeneous sources, employing a three-stage process and an adaptive query generator to access diverse knowledge types. The consistent performance improvements observed across different domains highlight the framework's effectiveness in enhancing LLM reliability and controllability.

Chapter 6 provided a comprehensive survey on retrieving multimodal information for augmented generation. This review synthesized methods for incorporating various modalities such as images, code, tables, graphs, and audio to enhance LLM capabilities. By offering insights into improving factuality, reasoning, interpretability, and robustness in multimodal AI systems, this survey contributes valuable knowledge to the rapidly evolving field of multimodal LLMs.

In conclusion, this thesis contributes advancements to the field of trustworthy and reliable NLP systems. By providing innovative solutions and comprehensive insights, we hope to foster the responsible and effective use of LLMs in real-life deployment and improve trust and confidence among users and stakeholders in the AI community. Collectively, these contributions advance our understanding of how to make LLMs more trustworthy, interpretable, and controllable without necessitating extensive model revisions or compromising performance. The frameworks and insights presented in this thesis offer practical solutions to longstanding challenges in LLM deployment, paving the way for more responsible and effective use of these powerful technologies across various domains.

7.2 Future Directions

As we look to the future, several promising directions emerge for further research. These include exploring the integration of our proposed frameworks, investigating

their applicability to emerging LLM architectures, and developing more approaches to address new challenges in multimodal and multi-task learning scenarios. Additionally, as the field of AI continues to evolve rapidly, there remains a critical need for ongoing research into ethical considerations, bias mitigation, and the development of robust evaluation metrics for trustworthy AI systems.

For the HI-Concept framework proposed in Chapter 3, there exist several promising avenues for future research and expansion.

- **User Interaction Studies:** Further research could focus on how users interact with the extracted features. This could involve developing intuitive interfaces for presenting High-Impact Concepts and conducting user studies to assess the effectiveness of these explanations in real-world scenarios. Such studies could provide valuable insights into how different user groups interpret and utilize these explanations, potentially leading to more tailored and effective explanation strategies.
- **Interactive Controllability:** Expanding beyond simple explanations, future work could explore enabling user interactions with the identified features to offer greater controllability over LLM predictions. This could involve developing mechanisms for users to adjust the influence of specific concepts, allowing for more fine-tuned control over model outputs. Such an approach could bridge the gap between explainability and controllability, offering users not just insights into model behavior but also the ability to guide it.
- **Extension to Text Generations:** Extending the HI-Concept framework to explain LLM generations, beyond simple classifications, represents a significant and complex challenge. This expansion would require adapting the methodology to handle the intricacies of sequence generation, potentially involving techniques to identify and explain concept impact across different parts of generated text. Such an extension could greatly enhance our understanding of LLM behavior in more complex, open-ended tasks.
- **Analyzing Large LLMs:** Applying the HI-Concept framework to study each layer of very large LLMs could offer valuable insights into their inner workings. Conducting such a comprehensive analysis could reveal how concepts are formed and transformed across layers, potentially uncovering

patterns or structures that are not apparent from studying the model as a whole. These insights could also offer inspirations for other fields such as model pruning, architecture design, or transfer learning.

The Verify-and-Edit (VE) framework presented in Chapter 4, followed by the Chain-of-Knowledge (CoK) framework shown in Chapter 5, also opens up several exciting avenues for future research:

- **Expanding Behavioral Modifications**: While the initial focus of VE was on improving factual correctness, the framework's potential extends far beyond this application. Future research could explore using CoT alterations for a variety of purposes:
  - Jailbreaking and Unreliability Detection: By systematically modifying CoTs, we could develop methods to identify vulnerabilities in LLMs, helping to improve their robustness against adversarial attacks.
  - Ethical and Safe Behavior Guidance: CoT alterations could be used to steer LLM outputs towards more ethical and safe behaviors. This could involve injecting ethical considerations into the reasoning process or redirecting potentially harmful lines of thought.
  - Bias Mitigation: Modifying CoTs could help in identifying and mitigating biases in LLM outputs, potentially leading to fairer and more inclusive language generation.
  - Creativity Enhancement: Alterations could be designed to encourage more creative or diverse thinking patterns in LLMs, potentially leading to more novel and varied outputs.
- **Multimodal Extensions:** Extending the VE framework to different modalities presents an exciting challenge and opportunity, such as vision-language models where VE could involve visual attention patterns or audio-text models where VE could help connect audio inputs to textual outputs. Furthermore, because of the reasoning component brought by CoTs, VE could be used to handle multi-modal reasoning effectively.
- **Real-time Editing and Interaction:** Developing methods for real-time CoT editing could enable more interactive and dynamic control over LLM

outputs. This could lead to systems where users can guide the reasoning process as it unfolds, offering unprecedented levels of control and customization.

Following our comprehensive survey of retrieving multimodal information for augmented generation, several promising directions for future research emerge beyond the ones already listed in Chapter 6:

- **Improving the reliability of vision-language models** The rapid advancement of vision-language models (VLLMs) has opened new frontiers in artificial intelligence, enabling sophisticated interactions between textual and visual modalities. However, like their text-only counterparts, VLLMs face challenges in reliability, particularly in terms of maintaining factual accuracy and reducing hallucinations. One promising approach to enhance the reliability of VLLMs is the incorporation of visual patch retrieval during the generation process. By dynamically retrieving and integrating relevant visual patches from a curated database, we can ground the model's outputs in concrete visual contexts, potentially mitigating hallucinations and improving factual consistency. This method could leverage the model's understanding of both textual and visual inputs to select the most pertinent visual information, thereby enriching the generation process with accurate, contextually relevant visual cues. Additionally, this approach could be extended to include multimodal knowledge bases, allowing the model to cross-reference textual and visual information for enhanced accuracy. As VLLMs continue to evolve and become more prevalent in various applications, developing such reliability-enhancing techniques will be crucial in addressing the current limitations of internal knowledge and reducing the frequency of hallucinations, ultimately leading to more trustworthy and effective vision-language AI systems.

- **Explainable Multimodal AI:** Developing methods for explaining the decisions and outputs of multimodal AI systems is crucial for building trust and understanding. Future research could focus on creating interpretable multimodal models and generating human-understandable explanations that incorporate multiple modalities. Choosing an appropriate and user-friendly unit of explanation would also be a unique challenge for multimodal models.

- **Robustness to Multimodal Adversarial Attacks:** As multimodal systems become more complex, ensuring their robustness to adversarial attacks

across different modalities becomes crucial. Future work should investigate techniques to detect and mitigate such attacks in multimodal contexts.

- **Ethical Considerations in Multimodal AI**: As multimodal AI systems become more prevalent, research into their ethical implications becomes crucial. Future work should address issues such as bias in multimodal datasets, privacy concerns in visual data, and the potential for misuse in deepfake generation.

Besides the frameworks and chapters in this thesis, there also exist several exciting venues for LLM trustworthiness and reliability in general:

- **Ethics and Bias Considerations:**
  - Comprehensive Bias Analysis: We can conduct studies to identify and quantify various forms of social, gender, racial, and cultural biases present in LLMs. This research should extend beyond surface-level outputs to examine biases in reasoning patterns and knowledge representation.
  - Bias Source Identification: Then, targeting these discovered biases, we can develop techniques to trace the origins of biases within LLM architectures, potentially identifying specific neurons or network components responsible for biased outputs. This could involve advanced interpretability methods and fine-grained analysis of the internals of the model.
  - Bias Mitigation Strategies: Finally, we should explore and compare different approaches to mitigate biases, such as curating new datasets for training, model architecture modifications, fine-tuning techniques, and post-processing methods.
- **Evaluating Trustworthiness and Reliability:** Currently, there is a lack of standardized metrics and datasets for systematically measuring the trustworthiness and reliability of LLMs. Creating a comprehensive evaluation framework that encompasses various aspects such as factual accuracy, consistency, bias, safety, and robustness would be of great benefit to the academic community. This framework should include a suite of benchmark datasets covering a wide range of domains, languages, and cultural contexts to ensure

a thorough evaluation of LLM performance and reliability. Additionally, designing and validating new metrics that can quantify different aspects of trustworthiness, such as measures for factual consistency, logical coherence, and alignment with human values, will be crucial. Research should also explore methods for continuous, real-time evaluation of LLM trustworthiness in deployment scenarios, allowing for adaptive trust management in practical applications.

- **Cross-cultural and Multilingual Trustworthiness:** As an important future extension, we can investigate how to ensure LLMs remain trustworthy and reliable across different cultural contexts, respecting diverse values and norms. Additionally, developing methods to maintain consistent levels of trustworthiness across multiple languages, addressing challenges related to low-resource languages and cultural nuances, will be crucial for the global applicability of these models.

List of Author's Awards, Patents, and Publications[1]

Conference Proceedings

- Han Cheol Moon, Shafiq Joty, **Ruochen Zhao**, Megh Thakkar, Xu Chi, Randomized Smoothing with Masked Inference for Adversarially Robust Text Classifications, in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (ACL 2023).

- **Ruochen Zhao***, Xingxuan Li*, Shafiq Joty, Chengwei Qin, 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) (ACL 2023).

- **Ruochen Zhao**, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, Shafiq Joty, Retrieving Multimodal Information for Augmented Generation: A Survey, in Findings of the Association for Computational Linguistics: EMNLP 2023.

- **Ruochen Zhao**, Shafiq Joty, Yongjie Wang, Tan Wang, Explaining Language Model Predictions with High-Impact Concepts, in Findings of the Association for Computational Linguistics: EACL 2024.

- Xingxuan Li*, **Ruochen Zhao***, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, Lidong Bing, Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources, in International Conference on Learning Representations (ICLR) 2024.

[1]The superscript * indicates joint first authors

- Bosheng Ding*, Chengwei Qin*, **Ruochen Zhao***, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, Shafiq Joty, Data augmentation using llms: Data perspectives, learning paradigms and challenges, in Findings of the Association for Computational Linguistics: ACL 2024.

Bibliography

[1] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022. xxii, 13, 55, 61, 64, 88, 98, 100

[2] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. xxii, 20, 56, 57, 59, 64, 65, 71, 84, 85, 88

[3] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. (arXiv:1706.07269), August 2018. doi: 10.48550/arXiv.1706.07269. URL `http://arxiv.org/abs/1706.07269`. arXiv:1706.07269 [cs]. 1

[4] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv:2004.03685 [cs]*, April 2020. URL `http://arxiv.org/abs/2004.03685`. arXiv: 2004.03685. 1

[5] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. 9

[6] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016. 9, 46

[7] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *stat*, 1050:2, 2017. 9, 10, 28, 33, 37

[8] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. 10

[9] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015. 10

[10] Robert C Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11:63–90, 1993. 10

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016. doi: 10.1145/2939672.2939778. 10, 11, 28

[12] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL `https://arxiv.org/abs/1705.07874`. 10

[13] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 11

[14] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017. 11, 12

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning, 2016. URL `https://arxiv.org/abs/1606.05386`. 11

[16] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018. 11, 12, 28

[17] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 12, 26, 31

[18] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020. 12, 28, 29, 33, 36, 37, 39, 41, 42

[19] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28, 2015. 13

[20] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. 13

[21] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019. 13

[22] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong,

China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1221. URL `https://aclanthology.org/D19-1221`. 13

[23] Danilo Croce, Daniele Rossini, and Roberto Basili. Auditing deep learning processes through kernel-based explanatory models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4037–4046, 2019. 13

[24] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018. 14

[25] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019. 14

[26] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017. 14

[27] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019. 15

[28] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR, 2019. 15

[29] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020. 15

[30] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2493–2500, 2020. 15

[31] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384. PMLR, 2019. 15

[32] Michael Harradon, Jeff Druce, and Brian Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018. 15

[33] Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Christopher Rozell, and Mark Davenport. Generative causal explanations of black-box classifiers. *Advances in Neural Information Processing Systems*, 33:5453–5467, 2020. 16

[34] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL `https://aclanthology.org/P18-1198`. 16

[35] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52, 2020. doi: 10.1162/coli_a_00367. URL `https://aclanthology.org/2020.cl-1.1`.

[36] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. 16

[37] Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, 2019. 16

[38] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. 16

[39] Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors. *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. 16

[40] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401, 2020. 16, 17

[41] Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *CoRR*, abs/2109.00725, 2021. URL `https://arxiv.org/abs/2109.00725`. 17

[42] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021. 17

[43] Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. Guided generation of cause and effect. In *Proceedings of*

the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-PRICAI-2020. International Joint Conferences on Artificial Intelligence Organization, July 2020. doi: 10.24963/ijcai.2020/502. URL `http://dx.doi.org/10.24963/ijcai.2020/502`. 17

[44] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*, 2019. 17

[45] David Alvarez-Melis and Tommi S Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*, 2017. 17

[46] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34, 2021. 17

[47] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*, 2021. 17

[48] Xi Ye and Greg Durrett. The unreliability of explanations in few-shot prompting for textual reasoning. 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=Bct2f8fRd8S`. 18, 20, 56, 61, 62, 63, 64, 66, 92

[49] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.47. URL `https://aclanthology.org/2022.naacl-main.47`. 18

[50] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 18

[51] 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. 18, 94, 96, 105

[52] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by

retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022. 18, 94

[53] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 19

[54] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model?, 2020. 19

[55] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL `https://aclanthology.org/N18-1074`. 19, 58, 62

[56] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. 19, 92

[57] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. True: Re-evaluating factual consistency evaluation, 2022. URL `https://arxiv.org/abs/2204.04991`. 19

[58] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021. 19

[59] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. *arXiv preprint arXiv:2107.06963*, 2021. 19

[60] Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*, 2022. 20

[61] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. 20, 61, 64, 70

[62] 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. 20

[63] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022. 20, 56

[64] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022. 20, 56, 61

[65] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*, 2020. 21

[66] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019. 21

[67] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models, 2020. URL `https://arxiv.org/abs/2004.12316`. 22

[68] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. URL `https://aclanthology.org/2021.eacl-main.24`. 22

[69] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018. 22

[70] Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. Content planning for neural story generation with aristotelian rescoring. *arXiv preprint arXiv:2009.09870*, 2020. 22

[71] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 22, 56, 91

[72] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021. 22

[73] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 22

[74] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 22, 56, 63, 91

[75] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR, 2023. 22

[76] 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. 22

[77] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`. 22

[78] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019. 23

[79] Irene Solaiman and Christy Dennison. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873, 2021. 23

[80] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL `https://arxiv.org/abs/2212.08073`. 23

[81] Sherin Mary Mathews. Explainable artificial intelligence applications in nlp, biomedical, and malware classification: a literature review. In *Intelligent computing-proceedings of the computing conference*, pages 1269–1292. Springer, 2019. 26

[82] Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. Discovering latent concepts learned in bert. In *International Conference on Learning Representations*, 2021. 26, 27, 33, 48, 51, 52

[83] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 27, 31, 34, 37

[84] Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17582–17596. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/701ec28790b29a5bc33832b7bdc4c3b6-Paper-Conference.pdf`. 27

[85] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33, 2020. 27

[86] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009. 27, 31

[87] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, June 2021. doi: 10.1162/coli_a_00404. URL `https://aclanthology.org/2021.cl-2.13`. 27, 33

[88] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791, 2021. 33, 35

[89] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 33

[90] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`. 34

[91] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015. 34

[92] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 34

[93] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 34

[94] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 35, 37, 41

[95] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005. 36

[96] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. 37, 41

[97] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(02)00060-2. URL `https://www.sciencedirect.com/science/article/pii/S0031320302000602`. Biometrics. 37, 41

[98] Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491. URL `https://aclanthology.org/2020.acl-main.491`. 46

[99] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.523. URL `https://aclanthology.org/2021.acl-long.523`. 46

[100] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977. 49

[101] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408, 2015. 50

[102] OpenAI-Blog. Chatgpt: Optimizing language models for dialogue, 2022. URL `https://openai.com/blog/chatgpt/`. 56, 67

[103] Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines. *arXiv preprint arXiv:2304.11076*, 2023. 56, 67

[104] Gary Marcus. Is chatgpt really a "code red" for google search?, 2022. URL `https://garymarcus.substack.com/p/is-chatgpt-really-a-code-red-for`. 56

[105] 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, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL `https://aclanthology.org/D18-1259`. 58, 62

[106] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL `https://aclanthology.org/2020.coling-main.580`. 58, 62

[107] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL `https://aclanthology.org/P17-1171`. 60, 65

[108] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 64, 91, 94

[109] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550. 65, 105

[110] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410. 65

[111] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023. 69, 105

[112] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. 70

[113] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513. 86, 87

[114] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *Proceedings of ICLR*, 2021. 86

[115] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. 91, 96

[116] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 91

[117] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 92

[118] OpenAI. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774. 94, 101

[119] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 92, 104

[120] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*, 2021. 92

[121] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*, 2023. 92

[122] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 92, 94

[123] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 93

[124] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864, 2020. 93

[125] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 105–124. Springer, 2022. 93

[126] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

[127] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[128] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 93

[129] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, volume 34, pages 13041–13049, 2020. 93

[130] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via

sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021.

[131] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 93

[132] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 93

[133] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

[134] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 93

[135] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 93

[136] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 88–105. Springer, 2022.

[137] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In *EMNLP*, pages 11238–11254. Association for Computational Linguistics, 2022.

[138] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *EMNLP*, pages 5558–5570. ACL, 2022. URL `https://aclanthology.org/2022.emnlp-main.375`. 93, 105

[139] Ruochen Zhao, Xingxuan Li, Yew Ken Chia, Bosheng Ding, and Lidong Bing. Can chatgpt-like generative models guarantee factual accuracy? on the mistakes of new generation search engines, 2023. 94

[140] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. Search engine guided neural machine translation. In *AAAI*, volume 32, 2018. 94

[141] Jason Weston, Emily Dinan, and Alexander Miller. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the*

2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI, pages 87–92, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5713. URL `https://aclanthology.org/W18-5713`. 94

[142] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1195. URL `https://aclanthology.org/D19-1195`. 94, 100

[143] Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1120. URL `https://aclanthology.org/N18-1120`. 94

[144] Jitao Xu, Josep-Maria Crego, and Jean Senellart. Boosting neural machine translation with similar translations. In *Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579. Association for Computational Linguistics, 2020.

[145] Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, 2021. 94

[146] Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. Response generation by context-aware prototype editing. In *AAAI*, volume 33, pages 7281–7288, 2019. 94

[147] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1124. URL `https://aclanthology.org/N19-1124`. 94

[148] Hao Peng, Ankur Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. Text generation with exemplar-based adaptive decoding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2555–2565, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1263. URL https://aclanthology.org/N19-1263. 94

[149] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74. 94

[150] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*, 2208, 2022. 94

[151] Hangfeng He, Hongming Zhang, and Dan Roth. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*, 2022. 94, 100

[152] 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*, 2022.

[153] Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. Verify-and-edit: A knowledge-enhanced chain-of-thought framework, 2023. 94

[154] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022. 96

[155] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520, 2022.

[156] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *CoRR*, abs/2206.10789, 2022.

[157] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022.

[158] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 96

[159] Weizhe Lin and Bill Byrne. Retrieval augmented visual question answering with outside knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11238–11254, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.772. URL `https://aclanthology.org/2022.emnlp-main.772`. 96

[160] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020. 96

[161] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, volume 36, pages 3081–3089, 2022. 96

[162] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.70. URL `https://aclanthology.org/2022.naacl-main.70`. 96

[163] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C.H. Hoi. Plug-and-play VQA: Zero-shot VQA by conjoining large pretrained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.67. URL `https://aclanthology.org/2022.findings-emnlp.67`. 96

[164] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society, 2017. 96

[165] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu

Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.375. URL https://aclanthology.org/2022.emnlp-main.375. 97

[166] Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. RAMM: retrieval-augmented biomedical visual question answering with multi-modal pre-training. *CoRR*, abs/2303.00534, 2023. 97

[167] Yucheng Zhou and Guodong Long. Style-aware contrastive learning for multi-style image captioning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2257–2267, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-eacl.169. 97

[168] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.39. URL https://aclanthology.org/2022.findings-naacl.39. 97

[169] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *CBMI*, pages 1–7. ACM, 2022. 97

[170] Zhan Shi, Hui Liu, Martin Renqiang Min, Christopher Malon, Li Erran Li, and Xiaodan Zhu. Retrieval, analogy, and composition: A framework for compositional generalization in image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1990–2000, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.171. URL https://aclanthology.org/2021.findings-emnlp.171. 97

[171] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.266. 97

[172] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Ming-Yu Liu, Yuke Zhu, Mohammad Shoeybi, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *CoRR*, abs/2302.04858, 2023. 97

[173] Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. Focus! relevant and sufficient context selection for news image captioning. In *Findings of*

the Association for Computational Linguistics: EMNLP 2022, pages 6078–6088, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.450. URL https://aclanthology.org/2022.findings-emnlp.450. 97

[174] Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyon Myaeng. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 897–906, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.113. URL https://aclanthology.org/2021.acl-short.113. 97

[175] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99, 2021. doi: 10.1162/tacl_a_00356. URL https://aclanthology.org/2021.tacl-1.6. 97

[176] Zujie Liang, Huang Hu, Can Xu, Chongyang Tao, Xiubo Geng, Yining Chen, Fan Liang, and Daxin Jiang. Maria: A visual experience powered conversational agent. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5596–5611, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.435. URL https://aclanthology.org/2021.acl-long.435. 97

[177] Lei Shen, Haolan Zhan, Xin Shen, Yonghao Song, and Xiaofang Zhao. Text is not enough: Integrating visual impressions into open-domain dialogue generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 4287–4296, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386517. doi: 10.1145/3474085.3475568. URL https://doi.org/10.1145/3474085.3475568. 97

[178] Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. Z-LaVI: Zero-shot language solver fueled by visual imagination. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1203, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.78. URL https://aclanthology.org/2022.emnlp-main.78. 97

[179] Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Wang, Miguel Eckstein, and William Yang Wang. Visualize before you write: Imagination-guided open-ended text generation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 78–92, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-eacl.5. 97

[180] Qingkai Fang and Yang Feng. Neural machine translation with phrase-level universal visual representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5687–5698, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.390. URL `https://aclanthology.org/2022.acl-long.390`. 97

[181] Xian Wu, Shuxin Yang, Zhaopeng Qiu, Shen Ge, Yangtian Yan, Xingwang Wu, Yefeng Zheng, S. Kevin Zhou, and Li Xiao. DeltaNet: Conditional medical report generation for COVID-19 diagnosis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2952–2961, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.261`. 97

[182] Simon Mille, Spyridon Symeonidis, Maria Rousi, Montserrat Marimon Felipe, Klearchos Stavrothanasopoulos, Petros Alvanitopoulos, Roberto Carlini Salguero, Jens Grivolla, Georgios Meditskos, Stefanos Vrochidis, and Leo Wanner. A case study of NLG from multimedia data sources: Generating architectural landmark descriptions. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 2–14, Dublin, Ireland (Virtual), 12 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.webnlg-1.1`. 97

[183] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 97

[184] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. *CoRR*, abs/2211.12561, 2022. 98

[185] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022. 98

[186] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022. 98

[187] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*, 2022. 98

[188] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas

Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 98

[189] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*, 2023. 98

[190] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *arXiv preprint arXiv:2301.13808*, 2023. 98

[191] Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Norman Rabe, Charles E Staats, Mateja Jamnik, and Christian Szegedy. Autoformalization with large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *NeurIPS*, 2022. URL `https://openreview.net/forum?id=IUikebJ1Bf0`. 99

[192] Makarius Wenzel, Lawrence C. Paulson, and Tobias Nipkow. The isabelle framework. In Ait Mohamed, Munoz, and Tahar, editors, *TPHOLs*, volume 5170 of *LNCS*, pages 33–38. Springer, 2008. 99

[193] Xixin Hu, Xuan Wu, Yiheng Shu, and Yuzhong Qu. Logical form generation via multi-task learning for complex question answering over knowledge bases. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1687–1696, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.145`. 99

[194] Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, Caiming Xiong, and Yingbo Zhou. Uni-parser: Unified semantic parser for question answering on knowledge base and database. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8858–8869, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.605. URL `https://aclanthology.org/2022.emnlp-main.605`. 99

[195] Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. TIARA: Multi-grained retrieval for robust question answering over large knowledge base. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8108–8121, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.555. URL `https://aclanthology.org/2022.emnlp-main.555`. 99

[196] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev.

FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022. doi: 10.1162/tacl_a_00446. URL https://aclanthology.org/2022.tacl-1.3. 99

[197] Feifei Pan, Mustafa Canim, Michael Glass, Alfio Gliozzo, and Peter Fox. CLTR: An end-to-end, transformer-based system for cell-level table retrieval and table question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 202–209, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.24. URL https://aclanthology.org/2021.acl-demo.24. 99

[198] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379, 2023. 99

[199] Weizhe Lin, Zhilin Wang, and Bill Byrne. FVQA 2.0: Introducing adversarial samples into fact-based visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 149–157, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-eacl.11. 99

[200] Kiran Ramnath, Leda Sari, Mark Hasegawa-Johnson, and Chang Yoo. Worldly wise (WoW) - cross-lingual knowledge fusion for fact-based visual spoken-question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1908–1919, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.153. URL https://aclanthology.org/2021.naacl-main.153. 99

[201] 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. 99

[202] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale seq2seq models to multi-document inputs. *arXiv preprint arXiv:1910.08435*, 2019. 99

[203] Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava. Retrieve and rerank: A simple and effective IR approach to simple question answering over knowledge graphs. In *Proceedings of the First Workshop on Fact Extraction*

and VERification (FEVER), pages 22–27, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5504. URL https://aclanthology.org/W18-5504. 99

[204] Soumajit Pramanik, Jesujoba Alabi, Rishiraj Saha Roy, and Gerhard Weikum. Uniqorn: unified question answering over rdf knowledge graphs and natural language text. *arXiv preprint arXiv:2108.08614*, 2021. 99

[205] Bowen Jin, Yu Zhang, Qi Zhu, and Jiawei Han. Heterformer: A transformer architecture for node representation learning on heterogeneous text-rich networks. *arXiv preprint arXiv:2205.10282*, 2022. 99

[206] Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. Fusing context into knowledge graph for commonsense question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1201–1207, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.102. URL https://aclanthology.org/2021.findings-acl.102. 99

[207] 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. 99

[208] Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. Empowering language models with knowledge graph reasoning for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9581, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.650. URL https://aclanthology.org/2022.emnlp-main.650. 99

[209] Qi Liu, Dani Yogatama, and Phil Blunsom. Relational memory-augmented language models. *Transactions of the Association for Computational Linguistics*, 10:555–572, 2022. doi: 10.1162/tacl_a_00476. URL https://aclanthology.org/2022.tacl-1.32. 99

[210] Chao-Hong Tan, Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, Can Xu, Huang Hu, Xiubo Geng, and Daxin Jiang. TegTok: Augmenting text generation via task-specific and open-world knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1597–1609, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.125. URL https://aclanthology.org/2022.findings-acl.125. 99

[211] Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5770–5793, Abu Dhabi, United

Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.390. URL https://aclanthology.org/2022.emnlp-main.390. 100

[212] Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5000–5009, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.387. URL https://aclanthology.org/2021.acl-long.387. 100

[213] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation, 2019. 100

[214] Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.238. URL https://aclanthology.org/2021.acl-long.238. 100

[215] Peter Jansen and Dmitry Ustalov. TextGraphs 2019 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5309. URL https://aclanthology.org/D19-5309. 100

[216] Xiaojing Yu and Anxiao Jiang. Expanding, retrieving and infilling: Diversifying cross-domain question generation with flexible templates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3202–3212, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.279. URL https://aclanthology.org/2021.eacl-main.279. 100

[217] Jia Xin, Wang Hao, Yin Dawei, and Wu Yunfang. Enhancing question generation with commonsense knowledge. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 976–987, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.87.

[218] Yunfan Gu, Yang Yuqiao, and Zhongyu Wei. Extract, transform and filling: A pipeline model for question paraphrasing based on template. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages

109–114, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5514. URL https://aclanthology.org/D19-5514. 100

[219] Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy X. R. Wang. D2S: Document-to-slide generation via query-based text summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1418, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.111. URL https://aclanthology.org/2021.naacl-main.111. 100

[220] Yixuan Su, Zaiqiao Meng, Simon Baker, and Nigel Collier. Few-shot table-to-text generation with prototype memory. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 910–917, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.77. URL https://aclanthology.org/2021.findings-emnlp.77. 100

[221] Yohan Jo, Haneul Yoo, JinYeong Bak, Alice Oh, Chris Reed, and Eduard Hovy. Knowledge-enhanced evidence retrieval for counterargument generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3074–3094, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.264. URL https://aclanthology.org/2021.findings-emnlp.264. 100

[222] Liying Cheng, Dekun Wu, Lidong Bing, Yan Zhang, Zhanming Jie, Wei Lu, and Luo Si. ENT-DESC: Entity description generation by exploring knowledge graph. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1187–1197, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.90. URL https://aclanthology.org/2020.emnlp-main.90. 100

[223] Keerthiram Murugesan, Mattia Atzeni, Pavan Kapanipathi, Kartik Talamadupula, Mrinmaya Sachan, and Murray Campbell. Efficient text-based reinforcement learning by jointly leveraging state and commonsense graph representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 719–725, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.91. URL https://aclanthology.org/2021.acl-short.91. 100

[224] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022. 100, 105

[225] Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Binding language models in symbolic languages. *ICLR*, 2023. 100

[226] Danilo Neves Ribeiro, Shen Wang, Xiaofei Ma, Rui Dong, Xiaokai Wei, Henghui Zhu, Xinchi Chen, Peng Xu, Zhiheng Huang, Andrew Arnold, and Dan Roth. Entailment tree explanations via iterative retrieval-generation reasoner. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 465–475, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.35. URL `https://aclanthology.org/2022.findings-naacl.35`. 100

[227] Zhicheng Yang, Jinghui Qin, Jiaqi Chen, Liang Lin, and Xiaodan Liang. LogicSolver: Towards interpretable math word problem solving with logical prompt-enhanced learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1–13, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.1. URL `https://aclanthology.org/2022.findings-emnlp.1`. 100

[228] 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, 2023. 100

[229] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5811–5820, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.515. URL `https://aclanthology.org/2020.acl-main.515`. 100

[230] Miaoran Li, Baolin Peng, Jianfeng Gao, and Zhu Zhang. Opera: Harmonizing task-oriented dialogs and information seeking experience. *arXiv preprint arXiv:2206.12449*, 2022.

[231] Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.41. URL `https://aclanthology.org/2022.findings-acl.41`.

[232] Silin Gao, Jena D. Hwang, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. ComFact: A benchmark for linking contextual commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1656–1675, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi:

10.18653/v1/2022.findings-emnlp.120. URL https://aclanthology.org/2022.findings-emnlp.120.

[233] Xing Han Lu, Siva Reddy, and Harm de Vries. The StatCan dialogue dataset: Retrieving data tables through conversations with genuine intents. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2799–2829, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.206. 100

[234] Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.15. URL https://aclanthology.org/2022.naacl-main.15. 100

[235] Fabian Galetzka, Jewgeni Rose, David Schlangen, and Jens Lehmann. Space efficient context encoding for non-task-oriented dialogue generation with graph attention transformer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7028–7041, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.546. URL https://aclanthology.org/2021.acl-long.546. 100

[236] Jia-Chen Gu, Zhenhua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. Filtering before iteratively referring for knowledge-grounded response selection in retrieval-based chatbots. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1412–1422, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.127. URL https://aclanthology.org/2020.findings-emnlp.127. 100

[237] Jaehun Jung, Bokyung Son, and Sungwon Lyu. AttnIO: Knowledge Graph Exploration with In-and-Out Attention Flow for Knowledge-Grounded Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3484–3497, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.280. URL https://aclanthology.org/2020.emnlp-main.280. 100

[238] Jun Zhang, Yan Yang, Chencai Chen, Liang He, and Zhou Yu. KERS: A knowledge-enhanced framework for recommendation dialog systems with multiple subgoals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1092–1101, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.94. URL https://aclanthology.org/2021.findings-emnlp.94. 100

[239] Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.168. URL https://aclanthology.org/2021.emnlp-main.168.

[240] Chieh-Yang Chen, Pei-Hsin Wang, Shih-Chieh Chang, Da-Cheng Juan, Wei Wei, and Jia-Yu Pan. AirConcierge: Generating task-oriented dialogue via efficient large-scale knowledge retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 884–897, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.79. URL https://aclanthology.org/2020.findings-emnlp.79. 100

[241] Sixing Wu, Ying Li, Dawei Zhang, and Zhonghai Wu. Improving knowledge-aware dialogue response generation by using human-written prototype dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1402–1411, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.126. URL https://aclanthology.org/2020.findings-emnlp.126. 100

[242] Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. Knowledge graph-augmented language models for knowledge-grounded dialogue generation, 2023. 101

[243] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022. 101

[244] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *arXiv preprint arXiv:2012.07331*, 2020. 101

[245] Jinming Zhao, Gholamreza Haffari, and Ehsan Shareghi. Generating synthetic speech from SpokenVocab for speech translation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1975–1981, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.findings-eacl.147. 101

[246] Minkyu Kim, Kim Sung-Bin, and Tae-Hyun Oh. Prefix tuning for automated audio captioning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 101

[247] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio:

Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023. 102

[248] Rafael Mestre, Stuart E. Middleton, Matt Ryan, Masood Gheasi, Timothy Norman, and Jiatong Zhu. Augmenting pre-trained language models with audio feature embedding for argumentation mining in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 274–288, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.findings-eacl.21`. 102

[249] Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 102

[250] Zihao He, Weituo Hao, and Xuchen Song. Recap: Retrieval augmented music captioner. *arXiv preprint arXiv:2212.10901*, 2022. 102

[251] Brandon Royal, Kien Hua, and Brenton Zhang. Deep composer: Deep neural hashing and retrieval approach to automatic music generation. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020. 102

[252] David M Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. Using external off-policy speech-to-text mappings in contextual end-to-end automated speech recognition. *arXiv preprint arXiv:2301.02736*, 2023. 102

[253] Alex Falcon, Giuseppe Serra, and Oswald Lanz. A feature-space multimodal data augmentation technique for text-video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4385–4394, 2022. 102

[254] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018. 102

[255] Shih-Hung Liu, Kuan-Yu Chen, Berlin Chen, Hsin-Min Wang, Hsu-Chun Yen, and Wen-Lian Hsu. Combining relevance language modeling and clarity measure for extractive speech summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(6):957–969, 2015. 102

[256] Benjamin Milde, Jonas Wacker, Stefan Radomski, Max Mühlhäuser, and Chris Biemann. Ambient search: A document retrieval system for speech streams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2082–2091, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/C16-1196`.

[257] Benjamin Milde, Jonas Wacker, Stefan Radomski, Max Mühlhäuser, and Chris Biemann. Demonstrating ambient search: Implicit document retrieval for speech streams. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 233–237, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL `https://aclanthology.org/C16-2049`. 102

[258] Tao Hu, Xuyu Xiang, Jiaohua Qin, and Yun Tan. Audio-text retrieval based on contrastive learning and collaborative attention mechanism. 2022. 102

[259] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. Audio-text retrieval in context. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4793–4797. IEEE, 2022.

[260] A Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*, 2022. 102

[261] Ramakanth Pasunuru and Mohit Bansal. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*, 2018. 103

[262] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.730. URL `https://aclanthology.org/2020.acl-main.730`. 103

[263] Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1859, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.145. URL `https://aclanthology.org/2020.emnlp-main.145`. 103

[264] Hung Le, Nancy Chen, and Steven Hoi. Vgnmn: Video-grounded neural module networks for video-grounded dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3377–3393, 2022. 103

[265] Xiang Long, Chuang Gan, and Gerard De Melo. Video captioning with multi-faceted attention. *Transactions of the Association for Computational Linguistics*, 6:173–184, 2018. 103

[266] Spencer Whitehead, Heng Ji, Mohit Bansal, Shih-Fu Chang, and Clare Voss. Incorporating background knowledge into video description generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3992–4001, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1433. URL `https://aclanthology.org/D18-1433`. 103

[267] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*, 2022. 103

[268] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *AAAI*, volume 29, 2015. 103

[269] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 103

[270] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. *CoRR*, abs/2302.14115, 2023. URL `https://doi.org/10.48550/arXiv.2302.14115`. 103

[271] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2VLAD: global-local sequence alignment for text-video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5079–5088. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00504. URL `https://openaccess.thecvf.com/content/CVPR2021/html/Wang_T2VLAD_Global-Local_Sequence_Alignment_for_Text-Video_Retrieval_CVPR_2021_paper.html`. 103

[272] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 103

[273] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 104

[274] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*, 2022. 104

[275] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 104

[276] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. *Mining of Massive Datasets, 2nd Ed.* Cambridge University Press, 2014. ISBN 978-1107077232. URL `http://www.mmds.org/`. 104

[277] Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.75. URL `https://aclanthology.org/2021.emnlp-main.75`. 105

[278] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL `https://aclanthology.org/2021.emnlp-main.552`. 105

[279] Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.496. URL `https://aclanthology.org/2021.emnlp-main.496`. 105

[280] Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. Learning dense representations of phrases at scale. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6634–6647, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.518. URL `https://aclanthology.org/2021.acl-long.518`. 105

[281] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL `http://proceedings.mlr.press/v139/radford21a.html`. 105

[282] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020. 105

[283] Emanuele Aiello, Lili Yu, Yixin Nie, Armen Aghajanyan, and Barlas Oguz. Jointly training large autoregressive multimodal models. *arXiv preprint arXiv:2309.15564*, 2023. 105

[284] Seyed Omid Davoudi and Majid Komeili. Toward faithful case-based reasoning through learning prototypes in a nearest neighbor-friendly space. In *International Conference on Learning Representations*, 2021. 105

[285] Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models, 2023. 106