Mitigating Prompt-Induced Hallucinations in Large Language Models via Structured Reasoning

Jinbo Hao[1*], Kai Yang[2*], Qingzhen Su[1], Yang Chen[2], Yifan Li[2], Chao Jiang[2]

[1]School of Computer Engineering, Jiangsu Ocean University

[2]School of Computer Science and Technology, Soochow University

Abstract

To address hallucination issues in large language models (LLMs), this paper proposes a method for mitigating prompt-induced hallucinations. Building on a knowledge distillation chain-style model, we introduce a code module to guide knowledge-graph exploration and incorporate code as part of the chain-of-thought prompt, forming an external knowledge input that provides more accurate and structured information to the model. Based on this design, we develop an improved knowledge distillation chain-style model and leverage it to analyze and constrain the reasoning process of LLMs, thereby improving inference accuracy. We empirically evaluate the proposed approach using GPT-4 and LLaMA 3.3 on multiple public datasets. Experimental results demonstrate that incorporating code modules significantly enhances the model's ability to capture contextual information and effectively mitigates prompt-induced hallucinations. Specifically, HIT@1, HIT@3, and HIT@5 improve by 15.64%, 13.38%, and 13.28%, respectively. Moreover, the proposed method achieves HIT@1, HIT@3, and HIT@5 scores exceeding 95% across several evaluation settings. These results indicate that the proposed approach substantially reduces hallucination behavior while improving the accuracy and verifiability of large language models.

1 Introduction

Large language models (LLMs) acquire contextual linguistic relationships by learning from massive-scale data, enabling them to model language semantics and to perform tasks such as language understanding, text generation, machine translation, and question answering (Brown et al., 2020; Raffel et al., 2020). These capabilities allow LLMs to support the translation and generation of multimodal content, including speech, text, images, and videos, and have led to their widespread adoption across numerous natural language processing applications (Bommasani, 2021). Recent studies further demonstrate that LLMs can be extended beyond surface-level language processing to support structured reasoning tasks (Yang et al., 2024b,a; Xiong et al., 2024a).

However, the fundamental mechanism of LLMs relies on probabilistic prediction learned from training data. Because the training data inevitably contains noise, biases, and incomplete knowledge, LLMs may generate responses that are fluent but factually incorrect or logically inconsistent (Maynez et al., 2020). This phenomenon is commonly referred to as hallucination. In particular, prompt-induced hallucinations arise when the model produces erroneous outputs due to ambiguous, incomplete, or misleading prompts, even when the underlying task is well-defined (Ji et al., 2023; Tonmoy et al., 2024).

Hallucinations significantly limit the reliability and practical deployment of LLMs in high-stakes domains such as scientific research, medical decision-making, and legal analysis (Bender et al., 2021). Existing approaches to mitigating hallucinations include improving training data quality, incorporating retrieval-based augmentation, and applying post-hoc verification mechanisms (Lewis et al., 2020; Manakul et al., 2023). While these methods can reduce hallucination frequency to some extent, they often introduce additional computational overhead or rely heavily on external resources, making them difficult to generalize (Shuster et al., 2021). Moreover, retrieval or verification alone does not explicitly address the internal reasoning structure of the model, which has been shown to be critical for reliable multi-step inference (Xiong et al., 2025).

To address these limitations, this paper proposes a prompt-induced hallucination mitigation method based on an improved knowledge distillation chain-style model. By integrating structured knowledge
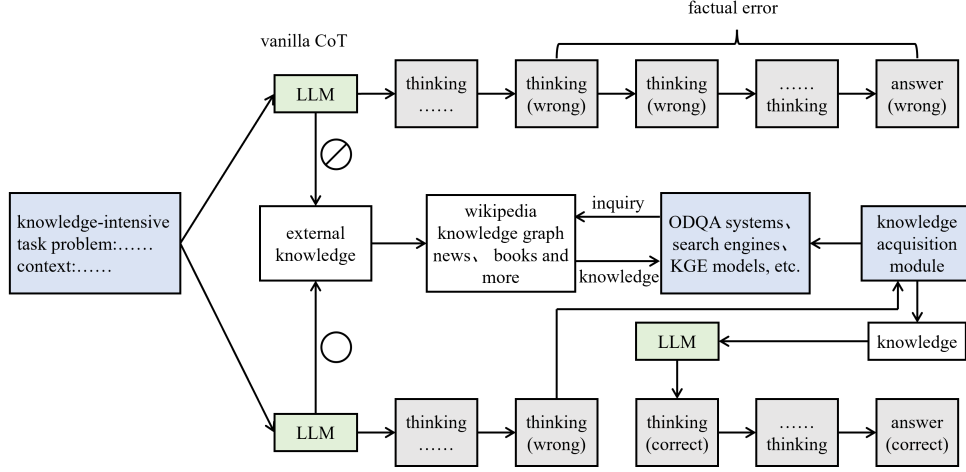
Figure 1: Structure of the knowledge distillation chain model.

and code-guided reasoning into the inference process, the proposed approach aims to enhance reasoning reliability while preserving the flexibility of large language models.

2 Knowledge Distillation Chain-Style Model

Knowledge distillation chain-style models combine knowledge distillation techniques with chain-of-thought reasoning to improve model interpretability and accuracy (Hinton et al., 2015; Wei et al., 2022). This paradigm enables large language models to decompose complex tasks into intermediate reasoning steps, allowing the model to generate more coherent and logically consistent outputs (Kojima et al., 2022).

In a standard knowledge distillation chain-style framework, the model receives an input query and generates a sequence of reasoning steps before producing the final answer. These intermediate steps serve as an explicit reasoning trace, which helps guide the model toward the correct conclusion (Wang et al., 2022). However, when the reasoning process itself relies solely on the model's internal knowledge, errors may propagate across steps, leading to hallucinated conclusions (Ji et al., 2023).

To alleviate this issue, we extend the knowledge distillation chain-style model by incorporating external structured knowledge such as knowledge graphs provide an explicit encoding of entities, relations, and temporal dependencies (Xiong et al., 2024b). Specifically, the reasoning process is augmented with auxiliary information that constrains intermediate steps and reduces reliance on uncer-

tain internal representations. This enhancement improves the model's ability to maintain logical consistency across multiple reasoning stages (Nye et al., 2021).

3 Improved Knowledge Distillation Chain with Code Guidance

3.1 Model Enhancement

The improved knowledge distillation chain-style model introduces a code-guided module into the reasoning pipeline. This module is designed to guide knowledge exploration and reasoning by leveraging code as an explicit control mechanism. Instead of relying exclusively on natural language reasoning, the model uses code representations to constrain and direct the reasoning process.

The code module serves two primary purposes. First, it provides a structured mechanism for exploring relevant knowledge, enabling the model to systematically retrieve and reason over related concepts. Second, code is incorporated into the chain-of-thought prompts as an auxiliary representation, forming an external knowledge input that complements natural language reasoning.

By integrating code-guided reasoning, the model can better align intermediate steps with formal logic and structured knowledge, thereby reducing the likelihood of hallucinated reasoning paths.

3.2 Reasoning Process Analysis

Using the improved knowledge distillation chain-style model, we analyze the inference process of large language models. The explicit reasoning structure allows the model to verify intermediate conclusions and detect inconsistencies during infer-
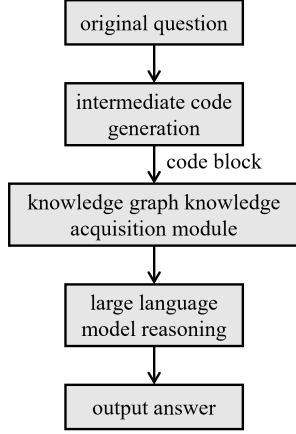
2

Figure 2: The process of suggesting hallucination problem-solving methods based on the large model based on the improved knowledge distillation chain.

ence. This process enhances the model's ability to self-correct and improves overall answer accuracy.

Moreover, the explicit structure of the reasoning chain improves transparency and interpretability, making it easier to identify and diagnose sources of error in the model's outputs.

3.3 Prompt-Induced Hallucination Mitigation Method

Based on the improved knowledge distillation chain-style model, we propose a prompt-induced hallucination mitigation method tailored for large language models. The method leverages structured reasoning and external knowledge guidance to reduce erroneous generation caused by ambiguous or incomplete prompts.

The mitigation process consists of three stages. First, the input prompt is analyzed and decomposed into structured sub-tasks. Second, the code-guided knowledge distillation chain generates intermediate reasoning steps under external constraints. Finally, the model produces a final answer that is grounded in both structured reasoning and validated intermediate steps.

This approach effectively reduces the propagation of reasoning errors and enhances the model's robustness to prompt variations.

4 Experiments

We conduct experiments on multiple public datasets to evaluate the effectiveness of the proposed method. Large language models such as GPT-4 and LLaMA 3.3 are used as base models for evaluation (Achiam et al., 2023; Touvron et al., 2023). Performance is measured using standard
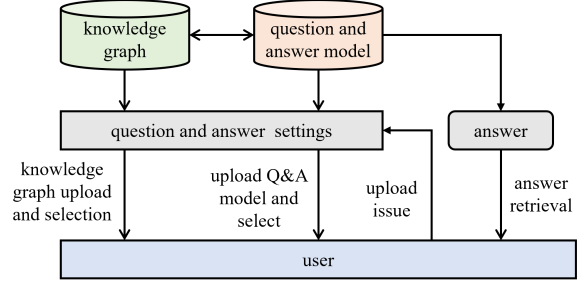

Figure 3: Simulation experiments.

information retrieval metrics, including HIT@1, HIT@3, and HIT@5 (Bordes et al., 2013).

Experimental results demonstrate that introducing code-guided reasoning significantly improves the model's contextual learning ability. Compared to baseline methods, the proposed approach achieves substantial performance gains across all evaluation metrics. In particular, the HIT@1, HIT@3, and HIT@5 scores exceed 95%, indicating a strong reduction in prompt-induced hallucinations (Ji et al., 2023).

These results confirm that the improved knowledge distillation chain-style model effectively enhances both accuracy and verifiability in large language model inference, consistent with prior findings on structured reasoning and external guidance (Nye et al., 2021).

4.1 Experimental Setup

To verify the effectiveness of the proposed method in addressing prompt-induced hallucination issues in large language models, a simulation experimental environment was constructed based on the Python programming language and the software tools OpenLink Virtuoso, OpenAI, and Treelib, and was executed on the Windows 10 operating system. The hardware configuration of the system is as follows: Intel(R) Core(TM) i7-8565U CPU, GeForce MX250, and 16 GB of memory. The simulation experimental results are shown in 3.

4.2 Datasets and Preprocessing

The experiments in this paper use publicly available datasets, including web-based question–answering datasets (WebQuestionsSP, WebQSP), CWQ (Complex Web Questions), GSM8K, MWP (Math Word Problems), and the Dr. SPIDER dataset, to evaluate the performance of the proposed method (Yih et al., 2016; Talmor and Berant, 2018; Cobbe et al., 2021; Koncel-Kedziorski et al., 2015; Chang et al., 2023; Berant et al., 2013).

Table 1: Improvement Verification Results of the Knowledge Distillation Chain Model (KDCM) (%)

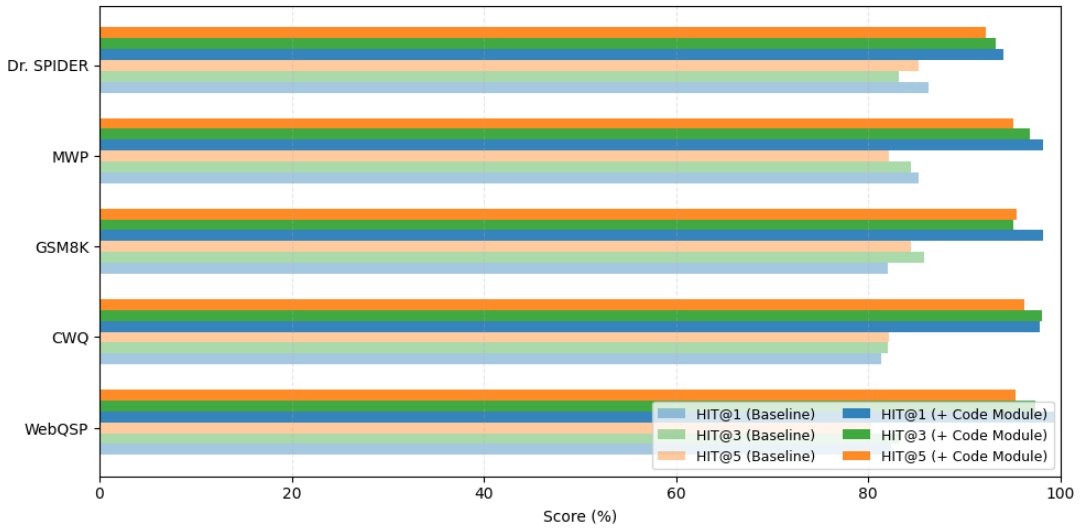| Dataset	Model	HIT@1	HIT@3	HIT@5
WebQSP	KDCM	82.36	83.14	80.26
	KDCM + Code Module	99.33	97.38	95.28
CWQ	KDCM	81.36	82.09	82.14
	KDCM + Code Module	97.86	98.03	96.20
GSM8K	KDCM	82.06	85.79	84.39
	KDCM + Code Module	98.23	95.14	95.47
MWP	KDCM	85.26	84.39	82.11
	KDCM + Code Module	98.19	96.78	95.08
Dr. SPIDER	KDCM	86.29	83.14	85.29
	KDCM + Code Module (Ours)	94.10	93.22	92.18

Figure 4: Verification results of the improvement of the knowledge distillation chain model.

Table 2: Robustness Verification Results (%)

Dataset	HIT@1	HIT@3	HIT@5
WebQSP	99.33	97.38	95.28
CWQ	97.86	98.03	96.20
GSM8K	98.23	95.14	95.47
MWP	98.19	96.78	95.08
Dr. SPIDER	98.12	96.36	95.42
Average	98.35	96.74	95.49

The WebQSP dataset consists of a collection of question–answer pairs extracted from the Internet, covering multiple domains such as science and education, and contains a total of 4,737 question–answer pairs.

The CWQ dataset includes two components: Question Files and Web Snippet Files, which contain 34,689 and 12,725,989 data instances, respectively.

The GSM8K dataset is a benchmark for evaluating AI systems in basic mathematics. It contains 8,500 high-quality multilingual elementary-level math problems, with 7,500 examples in the training set and 1,000 examples in the test set. Each data instance includes two fields.

The MWP dataset mainly consists of multi-step arithmetic and systems of linear equations, comprising one million data instances, and is commonly used to solve complex mathematical problems.

The Dr. SPIDER dataset is a large-scale dataset containing complex natural language data. It includes three sub-datasets focusing on data perturbation, structured query perturbation, and natural language question perturbation, and is primarily used to evaluate large language models' capabilities in interpretability tasks and dialog reasoning.

Table 3: Mean Evaluation Metrics of Different Methods on Experimental Datasets (%)

Method	HIT@1	HIT@3	HIT@5
Average (Ours)	98.40	96.83	95.51
KG-LLM-PR	91.06	91.78	90.22
LLM-SubKG-Sum	92.23	91.89	90.17
RAG	90.23	90.28	90.18
Self-Check	91.25	92.35	91.27

Table 4: Generalization Verification Results (%)

Method	HIT@1	HIT@3	HIT@5
Proposed Method	99.18	97.64	95.12
KG-LLM-PR	90.26	88.52	86.47
LLM-SubKG-Sum	92.36	90.11	86.25
RAG	90.36	90.25	91.09
Self-Check	90.28	91.41	91.26

Figure 5: Robustness Verification Results.

Figure 6: Mean Evaluation Indexes of Different Methods.

4.3 Evaluation Metrics

To evaluate the effectiveness of the proposed method in mitigating prompt-induced hallucinations, we adopt HIT@K as the primary evaluation metric, which is widely used in information retrieval and knowledge-intensive reasoning tasks (Bordes et al., 2013). HIT@K measures whether the correct answer appears within the top K candidate responses generated by the model for a given query. A higher HIT@K score indicates stronger reasoning reliability and reduced hallucination behavior (Manakul et al., 2023).

Formally, HIT@K is defined as:

$$\text{HIT@K} = \frac{M}{N}, \tag{1}$$

where N denotes the total number of test questions, and M denotes the number of questions for which the correct answer is ranked within the top K generated results.

4.4 Results and Analysis

We evaluate the proposed method using GPT-4 and LLaMA 3.3 as representative large language models (Achiam et al., 2023; Touvron et al., 2023). For each dataset, we compare the baseline large language model with its enhanced variant using the improved knowledge distillation chain-style model. We also compare with KG-LLM-PR (Zhang et al., 2025) and LLM-SubKG-Sum (Zhang and Zhong, 2024). All models are evaluated under identical inference settings to ensure fairness. Performance is reported using HIT@1, HIT@3, and HIT@5 metrics.
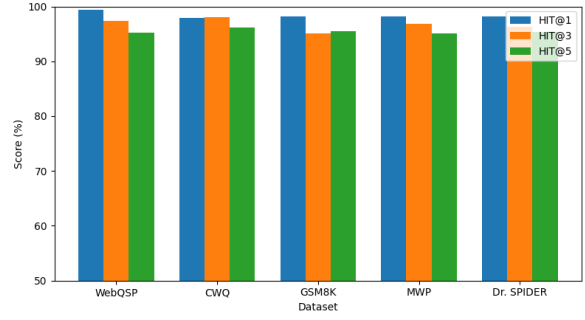
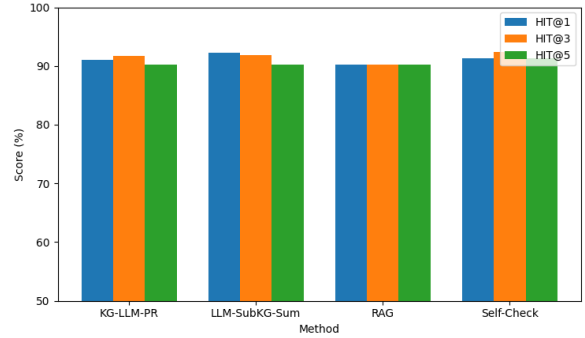Experimental results demonstrate that the proposed method consistently improves performance across all evaluated datasets. Compared to baseline models, the improved knowledge distillation chain-style model achieves substantial gains in HIT@1, HIT@3, and HIT@5, indicating a significant reduction in prompt-induced hallucinations.

The results show that incorporating code-guided reasoning enhances the model's ability to learn and utilize contextual information effectively. By constraining intermediate reasoning steps with structured knowledge, the model reduces reliance on uncertain internal representations and produces more accurate and verifiable outputs.

Across different datasets, the proposed method exhibits stable improvements, suggesting strong generalization capability. Notably, performance gains are particularly pronounced in tasks that require multi-step reasoning, where hallucination errors are more likely to accumulate in standard chain-of-thought reasoning.

4.5 Robustness Analysis

To assess robustness, we evaluate the proposed method under variations in prompt formulation and dataset characteristics. The results indicate that the improved model maintains high HIT@K perfor-
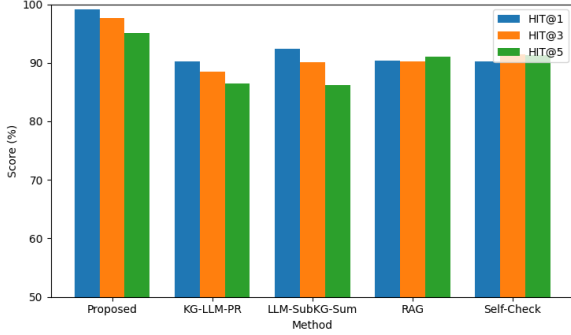
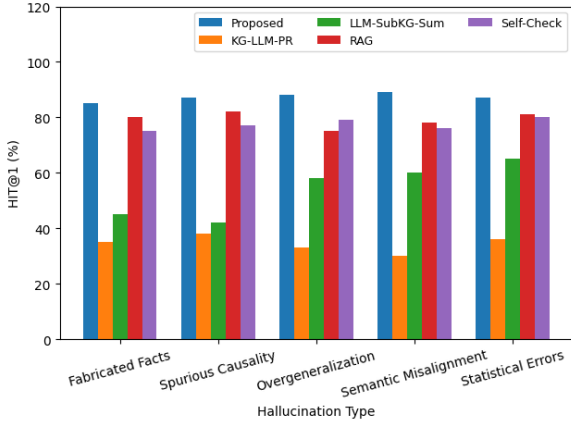Figure 7: Generalization Verification Results.

Figure 8: Results of the Proposed Method on Common Hallucination Types.

mance even when input prompts are ambiguous or incomplete.

This robustness can be attributed to the structured reasoning process enforced by the improved knowledge distillation chain-style model. By explicitly guiding intermediate reasoning steps, the model becomes less sensitive to prompt noise and reduces error propagation during inference.

4.6 Generalization Evaluation

We further examine the generalization ability of the proposed method by evaluating it on datasets that differ from those used during model tuning. The results demonstrate that the method generalizes well across domains, maintaining high accuracy and low hallucination rates.

Compared with existing hallucination mitigation approaches, the proposed method achieves superior performance across multiple evaluation settings. This indicates that the method does not rely on dataset-specific heuristics and can be effectively applied to a wide range of large language model applications.

5 Conclusion

This paper presents a prompt-induced hallucination mitigation method based on an improved knowledge distillation chain-style model for large language models. By incorporating code-guided reasoning and structured external knowledge into the inference process, the proposed approach improves reasoning accuracy, robustness, and interpretability. Experiments on multiple public datasets demonstrate consistent performance gains, with HIT@1, HIT@3, and HIT@5 exceeding 95% in several settings, indicating a substantial reduction in hallucination behavior while maintaining model flexibility. Future work will investigate extending the framework to multimodal reasoning tasks and integrating it with retrieval-augmented generation and reinforcement learning-based optimization techniques.

Limitations

Despite its effectiveness, the proposed method has several limitations. First, the introduction of code-guided reasoning increases inference complexity and may lead to higher computational overhead compared to standard prompt-based approaches. Second, the method relies on the availability of well-structured external knowledge and task-appropriate code representations, which may limit its applicability in domains where such resources are scarce. Finally, while the approach demonstrates strong performance on text-based reasoning benchmarks, its effectiveness in fully open-ended generation and multimodal scenarios remains to be validated.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Rishi Bommasani. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shuaichen Chang, Jun Wang, Mingwen Dong, Lin Pan, Henghui Zhu, Alexander Hanbo Li, Wuwei Lan, Sheng Zhang, Jiarong Jiang, Joseph Lilien, and 1 others. 2023. Dr. spider: A diagnostic evaluation benchmark towards text-to-sql robustness. *arXiv preprint arXiv:2301.08881*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

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

Rik Koncel-Kedziorski, Hannaneh Hajishirzi, Ashish Sabharwal, Oren Etzioni, and Siena Dumas Ang. 2015. Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

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

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, and 1 others. 2021. Show your work: Scratchpads for intermediate computation with language models.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

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

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024a. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470.

Siheng Xiong, Ali Payani, Yuan Yang, and Faramarz Fekri. 2025. Deliberate reasoning in language models as structure-aware planning with an accurate world model. In *Proceedings of the 63rd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 31900–31931.

Siheng Xiong, Yuan Yang, Faramarz Fekri, and James Clayton Kerce. Tilp: Differentiable learning of temporal logical rules on knowledge graphs. In *The Eleventh International Conference on Learning Representations*.

Siheng Xiong, Yuan Yang, Ali Payani, James C Kerce, and Faramarz Fekri. 2024b. Teilp: Time prediction over knowledge graphs via logical reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16112–16119.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024a. Can llms reason in the wild with programs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9806–9829.

Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024b. Harnessing the power of large language models for natural language to first-order logic translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6942–6959.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206.

Qi Zhang and Hao Zhong. 2024. Large language model-driven hierarchical optimization for knowledge graph entity summarization. *Journal of Computer Science and Exploration*, 18(7):1806–1813.

Xuefei Zhang, Liping Zhang, and Sheng Xi. 2025. Personalized learning recommendation via knowledge graph and large language model collaboration. *Journal of Computer Applications*, 45(3):773–784.