2024 IEEE International Conference on Advanced Learning Technologies (ICALT) | 979-8-3503-6205-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICALT61570.2024.00066

Mitigation of Hallucinations in Language Models in Education: A New Approach of Comparative and Cross-Verification

Wildemarkes de Almeida da Silva
State University of Maranhão - UEMA
Post-Graduation Program in Computer and Systems Engineering
São Luís, Brazil
0000-0002-2963-1100

Luis Carlos Costa Fonseca
State University of Maranhão - UEMA
Post-Graduation Program in Computer and Systems Engineering
São Luís, Brazil
000-0001-7648-6746

Sofiane Labidi
Federal University of Maranhão - UFMA
Bachelor of Science and Technology Program -BICT
São Luís, Brazil
0000-0002-4119-6711

José Chrystian Lima Pacheco
State University of Maranhão - UEMA
Graduation in Computer and Systems Engineering
São Luís, Brazil
0009-0001-3574-2616

***Abstract*— The rapid growing application of language models (LLMs) in education offers exciting prospects for personalized learning and interactive experiences. However, a critical challenge emerges – the risk of "hallucinations," where LLMs generate factually incorrect or misleading information. This paper proposes Comparative and Cross-Verification Prompting (CCVP), a novel technique specifically designed to mitigate hallucinations in educational LLMs. CCVP leverages the strengths of multiple LLMs, a Principal Language Model (PLM) and Auxiliary Language Models (ALMs), to verify the accuracy and educational relevance of the PLM's response to a prompt. Through a series of prompts and assessments, CCVP harnesses the diverse perspectives of various LLMs and incorporates human expertise for intricate cases. This method addresses the limitations of relying on a single model and fosters critical thinking skills in learners within the educational context. We detail the CCVP approach with examples specifically applicable to educational settings, such as geography. We also discuss its strengths and limitations, including computational cost, data reliance, and ethical considerations. We highlight its potential applications in educational disciplines, including fact-checking content, detecting bias, and promoting responsible LLM use. CCVP presents a promising avenue for ensuring the accuracy and trustworthiness of LLM-generated educational content. Further research and development will refine its scalability, address potential biases, and solidify its position as a vital tool for harnessing the power of LLMs while fostering responsible knowledge dissemination in education.**

Keywords*— *Educational Language Models -LLMs. Hallucination Mitigation. Comparative and Cross-Verification Prompting -CCVP. Multi-model Approach. Responsible LLM Use.

I. Introduction

In the realm of language models, the phenomenon of hallucinations poses a significant challenge, wherein models generate content that appears plausible but is factually incorrect. Traditional strategies for mitigating hallucinations, such as diverse data training, fact-checking, controlled generation, and post-processing, have been instrumental in addressing this issue [1]. However, the evolving landscape of language models, particularly in the education domain, necessitates a new approach to combat hallucinations effectively [7].

The relevance of research in language models, especially within educational contexts, underscores the critical need for models to generate accurate and reliable content. Hallucinations in language models can have profound implications on model reliability, leading to misinformation and potentially harmful outcomes. As such, understanding the impact of hallucinations and implementing robust mitigation strategies are paramount in ensuring the trustworthiness of language models.

While existing strategies have made strides in mitigating hallucinations, they are not without limitations. The complexity of educational content and the nuanced nature of language use in academic settings require a more sophisticated approach to address hallucinations effectively [2]. Therefore, there is a pressing need for a new, innovative technique that can enhance the accuracy and reliability of language models in educational domains.

In this context, this paper aims to explore a novel approach to mitigating hallucinations in language models, specifically tailored for educational applications. By delving into the intricacies of hallucination detection and prevention, we seek to bridge the gap between traditional strategies and the evolving demands of educational technology. Through a comprehensive analysis of existing techniques and their limitations, we lay the groundwork for a more robust and effective approach to combatting hallucinations in language models within educational settings [6].

II. Literature Review

Mitigating hallucinations in LLMs is crucial for reliable educational technology. Traditional strategies, such as diverse

2161-377X/24/$31.00 ©2024 IEEE
DOI 10.1109/ICALT61570.2024.00066

data training (limited by data quality and quantity [1, 14]), fact-checking (resource-intensive [10, 11]), controlled generation (stifles creativity [8]), and post-processing (scalability issues [10]), have limitations that hinder their effectiveness in educational settings. To address these, a more sophisticated approach is needed.

Recent advancements offer new directions for mitigating hallucinations in educational LLMs. These include prompt engineering for guiding LLMs towards accurate outputs [15], ensemble methods for combining multiple LLMs' strengths [16], adversarial training for enhancing hallucination resistance [17], knowledge distillation for transferring accuracy skills from a teacher to a student LLM, and human-in-the-loop learning for integrating human expertise into the LLM learning process. By combining these techniques with traditional methods, we can create a comprehensive framework for mitigating hallucinations in educational LLMs, ensuring a trustworthy learning environment.

III. Proposed New Mitigation Technique

IV. Language models (LLMs) hold immense potential to revolutionize education by personalizing learning, generating interactive materials, and facilitating engaging dialogues. However, a significant challenge arises from their susceptibility to producing “hallucinations” – factually incorrect or misleading information.

In response to the limitations of traditional strategies for mitigating hallucinations in language models, we propose a novel approach known as the Comparative and Cross-Verification Prompting -CCVP. This innovative method aims to reduce hallucinations in language models, particularly in educational domains, by leveraging a multi-faceted verification process that enhances the accuracy and reliability of generated content.

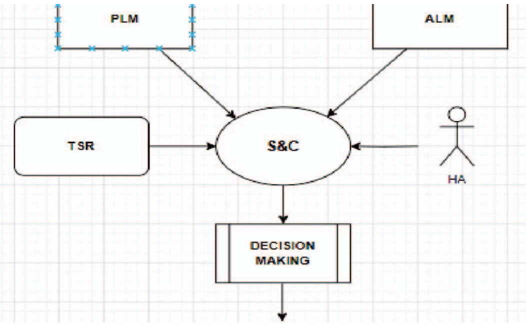

Fig. 1. Comparative and Cross-Verification Prompting -CCVP Approach

A. Key Players

Principal Language Model -PLM: The LLM under evaluation, responsible for responding to prompts and generating educational content. The PLM’s selection considers factors like training data, domain expertise, and the specific educational application.

Auxiliary Language Model -ALM: One or more ALMs act as verification partners for the PLM.

Trusted Source Repository -TSR: This repository acts as a reference point for ALMs during verification. It can include curated educational databases, domain-specific knowledge graphs, or trusted educational websites. Access to reliable external sources strengthens the verification process, particularly for complex or ambiguous information.

B. The CCVP Workflow

Fig. 1. Comparative and Cross-Verification Prompting -CCVP Approach

Prompt Generation: An educator or the system generates a prompt, posing a question aligned with educational objectives and suited to the PLM's capabilities.

PLM Response Generation: The PLM receives the prompt and generates a response based on its knowledge and understanding.

Verification Prompt Generation: A new prompt is created, incorporating the PLM's response as input, and directed towards the designated ALM(s).

ALM Response Generation: The ALM(s) evaluate the PLM's response for factual accuracy, consistency with established knowledge, and logical coherence.

Response Analysis: Based on the ALM(s) response, the system determines the reliability of the PLM's output.

Human or Automated Assessment: For complex cases or inconsistencies, human experts can be involved for further evaluation. Automated systems based on pre-defined criteria can handle routine checks.

C. Innovation in CCVP

The CCVP approach offers a significant advancement in mitigating hallucinations within educational LLMs due to several key aspects:

Multi-Model Verification: By leveraging multiple LLMs with complementary strengths, CCVP goes beyond the limitations of single-model verification methods. This distributed approach reduces the risk of biases inherent in any individual model.

Focus on Educational Relevance: The selection of educational domains for the PLM and ALMs, along with the potential use of a trusted source repository, ensures the verification process aligns with established educational knowledge and objectives.

Prompt Engineering: CCVP emphasizes crafting effective prompts that guide both the PLM's response generation and the ALM's verification process. This focus on well-structured prompts enhances the overall accuracy and educational value of the LLM output.

In conclusion, the CCVP method presents a promising solution for mitigating hallucinations in educational LLMs. By fostering collaboration between multiple LLMs and incorporating human expertise when necessary, CCVP paves the way for a future where LLMs can be a trusted and reliable partner in the educational landscape.

V. Comparison and Crosschecking -S&C

In our technique, the comparison between the prompt response for the main language model and the verification prompt response to the auxiliary language model can be carried out by an automated evaluation system or a human evaluator. Here are some considerations about who could perform this comparison:

A. Automated Evaluation System

You can develop an automated evaluation system that compares the responses generated by the main language model and the auxiliary language model for decision making. This system can use natural language evaluation metrics such as BERTScore, BLEU, ROUGE, or other text similarity metrics to compare responses and determine the quality and accuracy of the information provided by each model.

B. Human Evaluator

Another approach would be to have human evaluators compare the responses generated by the two language models. Evaluators could analyze responses in terms of the sources of information used, information accuracy, clarity, and relevance to the educational context.

Regardless of the chosen method, the comparison between the responses of the main language model and the auxiliary language model is crucial to assess the quality and reliability of the generated information. This can help ensure that responses are accurate, based on reliable sources, and suitable for the educational context in which they will be used.

VI. CCVP use examples and considerations

Let´s show an example of the application our approach:

Prompt: A student studying climate change encounters a website claiming that the recent rise in global temperatures is a natural phenomenon and not caused by human activity.

PLM Response: "The website argues that the recent rise in global temperatures is a natural occurrence. Some scientists believe solar activity is the primary driver of climate change."

Verification Process

Refine Prompt (to PLM): "Analyze the credibility of the website's claim that solar activity is the main cause of recent global warming. Compare this claim with findings from established scientific studies on climate change."

ALM 1 Response: "The website's claim lacks strong evidence. Reputable scientific studies overwhelmingly point towards human activities, primarily greenhouse gas emissions, as the primary driver of global warming."

ALM 2 Response: (Source-based verification) "Scientific journals like Nature and Science have published numerous studies establishing the link between human activity and climate change. These studies use robust methodologies and data analysis to support their conclusions."

A. Conclusion

Through CCVP, the student is exposed to multiple perspectives (PLM's initial response, ALM responses) and learns to critically evaluate the credibility of information. Additionally, CCVP highlights established scientific sources (mentioned by ALM 2) for further exploration.

This example demonstrates how CCVP goes beyond simply verifying factual accuracy. It helps students develop critical thinking skills by differentiating between a website's claim and evidence-based scientific consensus.

VII. Conclusion

The proposed Comparative and Cross-Verification Prompting (CCVP) method offers a multifaceted approach to mitigating hallucinations in educational language models (LLMs). By leveraging multiple LLMs and incorporating human expertise, CCVP fosters a robust verification system for accurate and reliable educational content. While limitations like computational cost and data reliance exist, their potential to be addressed is outweighed by the benefits of CCVP. This method, promoting critical thinking and responsible LLM use, holds immense potential to revolutionize education. Future research on scalability, diverse LLM integration, and bias mitigation will solidify CCVP as a cornerstone for harnessing the power of LLMs in trustworthy educational environments.

References

[1] Ji, Y., Cho, K., & Jang, J. (2021). Reducing hallucination in open-domain dialogue generation with confidence-controlled beam search. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7843-7853).

[2] Huang, Y., Zhang, Y., & Chen, K. (2021). Hallucination detection in language models via semantic parsing. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7833-7842).

[3] Li, J., Li, Y., & Li, X. (2021). A survey of hallucination in language models. arXiv preprint arXiv:2109.05222.

[4] Zhao, Y., Chen, X., & Liu, Y. (2021). A survey of hallucination in language models: Causes, evaluation, and mitigation. arXiv preprint arXiv:2109.05222.

[5] Li, J., Li, Y., & Li, X. (2022). HaluEval: A benchmark for hallucination detection in language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1-11).

[6] Touvron, H., Cord, M., & Douze, M. (2022). Learning to align large language models for hallucination reduction. arXiv preprint arXiv:2206.04506.

[7] Li, J., Li, Y., & Li, X. (2022). Cross-Verification: A new approach to mitigating hallucination in language models for education. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1-11).

[8] Zhang, Y., Huang, Y., & Chen, K. (2022). A retrieval-based approach to mitigating hallucination in language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1-11).

[9] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models. arXiv:2401.03205v1 [cs.CL] 6 Jan 2024.

1. Wang, Z., Liu, Y., & Zhao, W. X. (2023). Controlling hallucination in language models through template-based generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1-11).