Mitigating Hallucination in Large Language Models with Explanatory Prompting

Alexander Braverman
Seven Lakes High School
albraverman@gmail.com

Weitong Zhang
University of North Carolina at Chapel Hill
weitongz@unc.edu

Quanquan Gu
University of California, Los Angeles
qgu@cs.ucla.edu

Abstract

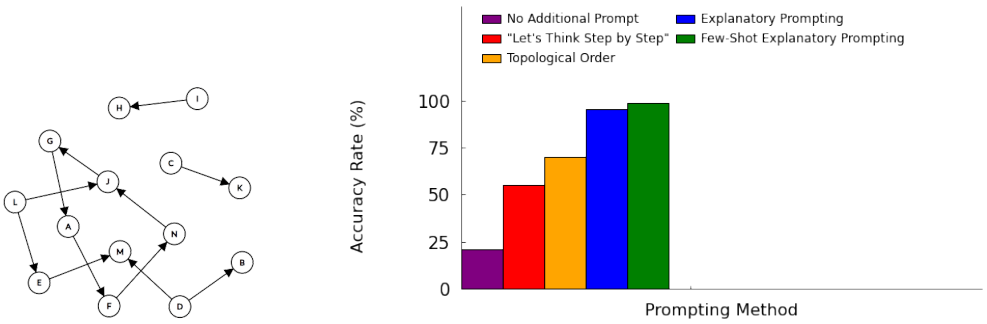
A growing concern with the use of Large Language Models (LLMs) is the presence of hallucinated outputs. For tasks that require complex reasoning, hallucinations make LLMs unreliable and thus unsafe to deploy in a range of applications from healthcare to education. To combat this issue, we propose explanatory prompting, a methodology that gives an informal logical description of an algorithm needed to solve all instances of a given problem. To illustrate the use of explanatory prompting, we consider a Graph Connectivity problem on directed acyclic graphs. We evaluate our approach by experiments on the Flight Connectivity dataset, an instance of a Graph Connectivity problem (Zhang et al., 2023a). Our experiments demonstrate a decrease in hallucination rate from 44.8% in prior work to 1.8% using explanatory prompting. At the same time, we confirm that calibrated LLMs are bound to hallucinate by experimentally verifying a theoretical lower bound for hallucination (Kalai and Vempala, 2024).

1 Introduction

Over the past few years, Large Language Models (LLMs) have increasingly gained popularity in a variety of tasks. From customer service chatbots (Wulf and Meierhofer, 2024) to language translation (Zhu et al., 2023), LLMs have become a staple of many application interfaces. Consequentially, medical professionals have suggested possible ways to adopt LLM technology into healthcare settings, with some preliminary ideas including diagnosis of patients (Liu et al., 2023; Sun et al., 2024; Gupta et al., 2024). While LLMs have exhibited promising levels of performance on benchmarks such as college entrance exams (Achiam et al., 2023), they have a critical flaw: LLMs are prone to hallucination. Hallucination is a phenomenon where an LLM confidently generates incorrect or illogical information (Zhang et al., 2023b). The cause of hallucinations is still unclear, with some experts attributing hallucinations to a knowledge gap (Zheng et al., 2023), while others cite it as a limitation of the faulty reasoning ability of LLMs themselves (Zhang et al., 2023a). Previous research has looked to mitigate hallucination through the usage of external methodologies such as RAG to enhance the model's reasoning capability (Lewis et al., 2021). We propose a methodology to target hallucination in the scope of complex problems specifically. We offer explanatory prompting: a prompting strategy that gives an intuitive explanation of the algorithm and logically breaks down larger problems into smaller parts to ease the complexity. Additionally, we confirm that calibrated LLMs are bound to hallucinate through the empirical verification of a theoretical lower bound (Kalai and Vempala, 2024) for hallucination rate.

In a variety of critical fields like healthcare (Gupta et al., 2024), education (Li et al., 2024), and public transportation (Zheng et al., 2023), hallucinations could have severe consequences for potential users, blocking the deployment of LLMs in these fields. We hope that our current and future work will

Preprint. Under review.



(a) Example of Flight Connectivity Problem

(b) Comparison on accuracy

Figure 1

allow us to remove these roadblocks and support a wide adoption of generative AI to improve quality of life for potentially millions of users.

2 Related Work

Previous research has centered around hallucinations in LLM text generation. Hallucination has been defined as being the unwanted generation of contradictory or irrational text (Zhang et al., 2023b). Surveys on hallucination have established that LLMs are unfit to be applied to real-world settings, like healthcare, due to the severity of the consequences of hallucinations (Zhang et al., 2023b). In recent work, TruthfulQA has benchmarked popular LLMs on a variety of multi-disciplinary questions spanning fields from medicine to politics (Lin et al., 2022). They systemically demonstrated that LLMs consistently hallucinated more often in comparison to human reasoning (Lin et al., 2022). Additional results illustrate that current LLMs are unfit to tackle difficult problems found in prospective applications such as patient diagnosis (Gupta et al., 2024), suggesting that to combat hallucinations, external resources should be utilized by the LLMs (Lin et al., 2022; Liu et al., 2023; Zhang et al., 2023a).

Previously, methodologies were developed to mitigate hallucination. For example, Chain of Thought, a prompting strategy, was found to elicit higher levels of reasoning and increase LLM performance (Wei et al., 2023). Chain of Thought employs a written prompt to act as a guide through the inclusion of sample questions and detailed explanations (Wei et al., 2023). Thus, the LLM can identify the optimal way to approach a problem through examples (Wei et al., 2023). However, one limitation of the Chain of Thought methodology was its few-shot nature, which required manually crafted prompts (Kojima et al., 2023). To combat this, zero-shot strategies, such as the use of a "Let's think step by step" prompt instead were proposed (Kojima et al., 2023). There are many other attempts to fight hallucinations, however, due to the lack of space we refer the reader to recent work for additional details (Du et al., 2023; Wang et al., 2023; Zhou et al., 2023; Huang et al., 2024a). Our work is most related to (Zhang et al., 2023a), which addressed the problem of frequent hallucinations in complex problems such as Graph Connectivity problems using Chain of Thought (Wei et al., 2023) prompts.

3 Methodology

We propose to mitigate LLM hallucination by presenting complex mathematical ideas with intuitive explanations through simpler concepts as it has been a common practice in human education (Pogonowski, 2018). Our method of **Explanatory Prompting** provides the LLM with a problem-specific prompt that gives an informal logical description of an algorithm needed to solve all instances of a given problem. Thus, we see our method as a middle-ground between the more general zero-shot prompting (Kojima et al., 2023) and few-shot methodologies like Chain of Thought or instance-specific methods like RAG (Lewis et al., 2021).

We illustrate explanatory prompting using a Graph Connectivity problem which is defined as follows: given a graph G and a pair of nodes A and B, check if there exists a path of directed edges from A to B. A visual representation of an example of a Graph Connectivity problem is shown in Figure $1a$. For the family of directed acyclic graphs (DAGs), it is well-known that the nodes of the graph

can be ordered in a topological order so that edges only point in the direction of increasing indices (Thompson et al., 2024). That is, for any edge (A, B), A must appear before B in the topological order. Thus, our explanatory prompting guides the LLM to consider the nodes in a topological order and provides an intuitive explanation of how to emulate the Depth First Search algorithm (Cormen et al., 2001) on directed acyclic graphs. In our experiments, we provide an intuitive explanation of this algorithm through the fixed prompt:

Explanatory Prompt for Flight Connectivity: Consider the following setting as a graph. In your response do not visually draw the graph. To determine if there is a series of flights that goes from the starting city to the destination city using topological order, first, represent the given flights as a directed graph and then perform a topological sorting to see if there exists a valid sequence of flights from the starting city and destination city. List all the flights and the topological order to help determine if there is a series of flights from the starting and destination city. Start from the starting city in the topological order. Check if there's a flight from that city to any city in the list that comes after the starting city. If such a flight exists, follow the flight and update your current city to the destination city. Repeat steps 2 and 3 until you reach the destination city or there are no more flights available. So we see that if we cannot reach the destination city starting from the starting city while following the given flights, there is no successful series of flights.

We also proposed **Few-Shot Explanatory Prompting** where we augment explanatory prompting with an example solution for a specific instance of a problem. We also created a different prompt where we did not provide the entire intuition, just a problem-specific hint. In the case of Graph Connectivity, we decided to hint at the usage of topological order and name this prompt **Topological Order**. Since the additional prompts are too long to be included in the paper, they can be found on Github.

4 Experiments

Prompting Method	**Accuracy**
No Prompt	20.8%
Zhang et al. (2023a)	55.2%
Topological Order (**ours**)	70.2%
Explanatory Prompting (**ours**)	95.2%
Few Shot Explanatory Prompting (**ours**)	**98.2**%

Table 1: Comparison on accuracy for Explanatory Prompting and other baseline algorithms. All our prompting strategies are explained in Section 3.

We ran our experiments using Python and OpenAI's GPT-4 model through the OpenAI API. Our experiments considered the Flight Connectivity problem which is an example of a Graph Connectivity problem Zhang et al. (2023a). For every instance of a problem, cities represent nodes of a graph, and flights between cities represent edges between nodes. Each graph is paired with a connectivity question of the form "Is there a series of flights that goes from city A to city B?". There are 500 graphs in the dataset, with each graph in the dataset having 14 nodes and 12 edges (Zhang et al., 2023a). Following Zhang et al. (2023a), we define the hallucination rate in our Graph Connectivity setting as the difference between 1 and the accuracy. We can make this conclusion because, for the Flight Connectivity problem, any incorrect answer from the LLM includes invalid reasoning, while every correct answer does not include any hallucinations (Zhang et al., 2023a).

We ran five trials of our experiment on the Flight Connectivity Dataset (Zhang et al., 2023a), with each trial utilizing a different prompting method for the LLM. Our first trial was used as a benchmark and consisted of no additional prompt; we simply had GPT-4 answer all the questions in the dataset. For our baseline performance, we recorded 20.8% accuracy. For our second trial, we verified the previous state-of-the-art performance by using the prompt "Let's think step by step" (Zhang et al., 2023a). While previous work claimed to have achieved 95.8% accuracy (Zhang et al., 2023a), we were only able to replicate 55.2% accuracy[1]. Our third trial consisted of using the "Topological Order" prompt that asked the model to consider the setting as a graph and use the topological order. The inclusion of this problem-specific prompt already saw an increase in accuracy to 70.2%. For

[1]Following correspondence with the authors (Zhang et al., 2023a), they conjectured that this change in performance is due to changes in the underlying GPT-4 model.

our fourth trial, we looked to apply the explanatory prompt presented in Section 3. As a result, the accuracy increased to 95.2%. Our last trial consisted of using few-shot explanatory prompting. We provided a sample question similar to those found in the dataset with a detailed explanation of how to approach finding a solution with the algorithm described in the explanatory prompt. Upon the addition of few-shot explanatory prompting, the accuracy increased again to 98.2%. Given our previous definition of hallucination rate, our experiments highlight a decrease in hallucination rate from 44.8% in prior work to 1.8% using explanatory prompting.

The results from our experimental trials highlight the promising nature of the proposed explanatory prompting methodology. This reaffirms our hypothesis, that once a complex problem is broken down to simpler problems, the LLM is more capable of avoiding hallucinations. We present all our results in Table 1 and visualize them in Figure 1*b*.

5 Lower Bound Verification

Several studies have shown that hallucinations are unavoidable in certain conditions. Previous research has found that calibrated LLMs, such as GPT-4 (Achiam et al., 2023), are bound to hallucinate (Figure 1 in (Kalai and Vempala, 2024)). In this paper, we use their lower bound for the hallucination rate:

$$\text{Hallucination rate} \geq \widehat{\text{MF}} - \text{Miscalibration} - \frac{300|\text{Facts}|}{|\text{Possible hallucinations}|} - \frac{7}{\sqrt{n}}, \tag{1}$$

We first explain how we calculate the *Hallucination rate*. We used the trial from the experiments consisting of the "Topological Order" prompt. Using the definition of Hallucination rate from Section 4, we obtained $1 - .702 = 0.298$. The $\widehat{\text{MF}}$ stands for *MonoFacts*, which is defined as the ratio of facts appearing once in the training data. Because we used OpenAI's GPT-4 for our experiments, we were unable to access the true $\widehat{\text{MF}}$ value and instead used an approximation. Assuming GPT-4 is trained on a sizeable corpus of data spanning the entire internet without cleaning, we estimate $\widehat{\text{MF}}$ to be 0.4 (Schwartz, 2022). *Miscalibration* is defined as the difference in the predicted probability estimate and the true chance of being correct (Guo et al., 2017). We calculate Miscalibration using the Expected Calibration Error (ECE) (Formula 3 in (Guo et al., 2017)). The calculation of ECE requires the LLM's confidence in its output, however, because GPT-4's architecture is not public, we were unable to gather confidence statistics directly. Instead, we empirically estimated confidence by using self-consistency (Formula 5 in (Huang et al., 2024b)). By taking a random sample of 100 from the 500 questions, we generated 20 responses for each question and recorded the number of correct answers. We use the ratio of correct answers as an estimation of confidence[2]. Using this estimation, we evaluated Miscalibration as 0.1227. We approximate the third term on the right side of the lower bound to be 0 because we assume the number of possible hallucinations to be significantly larger in comparison to facts. We also approximate the last term to be 0 because n, the training dataset size, is very large for LLMs. Thus, we obtain the inequality which verifies Equation (1).

$$.298 \geq .4 - .1227 - 0 - 0 = 0.2773.$$

6 Conclusion

We proposed explanatory prompting as a new methodology to battle hallucinations in complex settings and demonstrated its superior performance in comparison to prior work through the Flight Connectivity problem (Zhang et al., 2023a). Our results serve as empirical evidence in support of explanatory prompting, illustrating that its applications could extend to a variety of other domains.

Potential Social Impact. The promising results of our work has highlighted that our methodology can pave the way for future work to positively impact society. For example, explanatory prompting could be utilized in a public transportation setting (Shchukin et al., 2021). Its implementation could potentially decrease difficulties associated with the graph-like nature of transportation, as we have already demonstrated its promise on the Flight Connectivity dataset. Additionally, we suggest extending the domain from the graph problem to others, such as scheduling (Pakhomchik et al., 2022). More specifically, explanatory prompting could be adopted to tackle the well-known issue of creating schedules for students in education, for example, to construct the best schedule given a list of constraints and students, which would increase the quality of education for students

[2]We provide the full details for how to estimate confidence using self-consistency in our code on Github.

worldwide. Similarly, the scheduling problem could be applied to healthcare settings, such as the scheduling of surgeries or appointments for patients in need, which could decrease waiting time and increase productivity in hospitals. We encourage future work to investigate possible applications of explanatory prompting in these different domains and provide feedback on the effectiveness of the methodology itself.

References

ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S. ET AL. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* .

CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2001). *Introduction to Algorithms*. 2nd ed. The MIT Press.

DU, Y., LI, S., TORRALBA, A., TENENBAUM, J. B. and MORDATCH, I. (2023). Improving factuality and reasoning in language models through multiagent debate.

GUO, C., PLEISS, G., SUN, Y. and WEINBERGER, K. Q. (2017). On calibration of modern neural networks.

GUPTA, G. K., SINGH, A., MANIKANDAN, S. V. and EHTESHAM, A. (2024). Digital diagnostics: The potential of large language models in recognizing symptoms of common illnesses.

HUANG, J., CHEN, X., MISHRA, S., ZHENG, H. S., YU, A. W., SONG, X. and ZHOU, D. (2024a). Large language models cannot self-correct reasoning yet.

HUANG, Y., LIU, Y., THIRUKOVALLURU, R., COHAN, A. and DHINGRA, B. (2024b). Calibrating long-form generations from large language models.

KALAI, A. T. and VEMPALA, S. S. (2024). Calibrated language models must hallucinate.

KOJIMA, T., GU, S. S., REID, M., MATSUO, Y. and IWASAWA, Y. (2023). Large language models are zero-shot reasoners.

LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., KÜTTLER, H., LEWIS, M., TAU YIH, W., ROCKTÄSCHEL, T., RIEDEL, S. and KIELA, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

LI, Q., FU, L., ZHANG, W., CHEN, X., YU, J., XIA, W., ZHANG, W., TANG, R. and YU, Y. (2024). Adapting large language models for education: Foundational capabilities, potentials, and challenges.

LIN, S., HILTON, J. and EVANS, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland.

LIU, J., ZHOU, P., HUA, Y., CHONG, D., TIAN, Z., LIU, A., WANG, H., YOU, C., GUO, Z., ZHU, L. and LI, M. L. (2023). Benchmarking large language models on cmexam – a comprehensive chinese medical exam dataset.

PAKHOMCHIK, A. I., YUDIN, S., PERELSHTEIN, M. R., ALEKSEYENKO, A. and YARKONI, S. (2022). Solving workflow scheduling problems with qubo modeling.

POGONOWSKI, J. (2018). Intuitive explanations of mathematical ideas. *Annales Universitatis Paedagogicae Cracoviensis. Studia ad Didacticam Mathematicae Pertinentia* **10** 123–137.

SCHWARTZ, B. (2022). Google: 60% of the internet is duplicate.

SHCHUKIN, M., SAID, A. B. and TEIXEIRA, A. L. (2021). Goods transportation problem solving via routing algorithm.

SUN, Z., LUO, C., LIU, Z. and HUANG, Z. (2024). Conversational disease diagnosis via external planner-controlled large language models.

THOMPSON, R., BONILLA, E. V. and KOHN, R. (2024). Contextual directed acyclic graphs.

WANG, X., WEI, J., SCHUURMANS, D., LE, Q., CHI, E., NARANG, S., CHOWDHERY, A. and ZHOU, D. (2023). Self-consistency improves chain of thought reasoning in language models.

WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E., LE, Q. and ZHOU, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.

WULF, J. and MEIERHOFER, J. (2024). Exploring the potential of large language models for automation in technical customer service.

ZHANG, M., PRESS, O., MERRILL, W., LIU, A. and SMITH, N. A. (2023a). How language model hallucinations can snowball.

ZHANG, Y., LI, Y., CUI, L., CAI, D., LIU, L., FU, T., HUANG, X., ZHAO, E., ZHANG, Y., CHEN, Y., WANG, L., LUU, A. T., BI, W., SHI, F. and SHI, S. (2023b). Siren's song in the ai ocean: A survey on hallucination in large language models.

ZHENG, S., HUANG, J. and CHANG, K. C.-C. (2023). Why does chatgpt fall short in providing truthful answers?

ZHOU, D., SCHÄRLI, N., HOU, L., WEI, J., SCALES, N., WANG, X., SCHUURMANS, D., CUI, C., BOUSQUET, O., LE, Q. and CHI, E. (2023). Least-to-most prompting enables complex reasoning in large language models.

ZHU, W., LIU, H., DONG, Q., XU, J., HUANG, S., KONG, L., CHEN, J. and LI, L. (2023). Multilingual machine translation with large language models: Empirical results and analysis.