HalluAttack: Mitigating Hallucinations in LLMs via Counterfactual Instruction Fine Tuning

Anonymous EMNLP submission

Abstract

LLMs encapsulate a vast range of world knowledge with huge mount of pretraining data. While these models have demonstrated remarkable capabilities in various applications, they are prone to generating content infused with hallucinations, compromising the trustworthiness of their output. This phenomenon raises concerns of LLM applications, particularly when the dissemination of misleading information can have detrimental impacts. In this paper, we propose a simple yet effective method called HalluAttack which generates high quality counterfactual instruction data in order to reduce the hallucinations. We observe that these counterfactual instruction data can unlock the self-reflection ability of LLMs, and the LLMs will use knowledge learnt from pretraining phase more accurately. We conducted experiments across multiple open-source LLMs to evaluate the effectiveness of our proposed approach[1]. Results consistently demonstrate that, through counterfactual attack and subsequent fine-tuning, we are able to significantly improve the model performance on hallucination benchmarks (e.g. TruthfulQA and HaluQA). Moreover, we also find that the LLMs fine-tuned with counterfactual instruction data can also achieve gains on public general benchmarks like C-Eval, MMLU and GSM8K, which also demonstrate the effectiveness of our approach on hallucination mitigation.

1 Introduction

Recently, the advent of large language models (LLMs) has shown unprecedented levels of performance across a myriad of NLP tasks. These models, such as GPT-4(Achiam et al., 2023), LLaMA(Touvron et al., 2023) and QWen(Bai et al., 2023), etc, trained on extensive corpora, have exhibited remarkable abilities to generate coherent and contextually relevant text. However, an emerging concern is the propensity for these models to "hallucinate", producing text that, while fluent, is factually incorrect or entirely fabricated(Ji et al., 2022). This tendency not only undermines the credibility of model outputs but also poses significant risks in applications requiring high levels of accuracy and reliability, such as in financial, medical or legal area.

Due to the importance of understanding the factuality and hallucination of LLMs, there have been substantial research interest from academic community(Liu et al., 2024; Tonmoy et al., 2024; Li et al., 2024; Luo et al., 2024; Huang et al., 2023a; Sun et al., 2024). One of the most common approach to mitigate hallucination of LLMs is Retrieval Augmented Generation(RAG)(Lewis et al., 2020; Guu et al., 2020; Shuster et al., 2021; Shi et al., 2023b; Yu et al., 2022; Luo et al., 2023). This method leverages relevant documents retrieved from an external knowledge source to enhance the generation process. However, introducing an external knowledge base and a complex retrieval system is cost,

Figure 1: Example of hallucination with counterfactual prompt[2]

[1] The data we used for fine-tuning is publicly available in https://github.com/oldstree/halluattack

[2] Generated by Qwen1.5-32B-Chat

and actually it doesn't eliminate the intrinsic hallucinations of LLMs themselves. Another common approach to mitigate hallucination of LLMs is to enhance the factual correctness of the training data. A notable example is phi(Gunasekar et al., 2023) which uses a section of "textbook quality" data from the web during the pretraining phase. This kind of approach can only be used when we want to train a LLM from scratch. However, the huge amount the training data and large number of parameters of LLMs presents significant challenges and high costs to retrain a LLM. Knowledge editing(Cao et al., 2021; Yao et al., 2023; Tian et al., 2024) recently attracts research interests from researchers. It fixes factual errors by editing some specific "neurons" in LLMs. While knowledge editing can effectively mitigate the model's knowledge gap to some extent, it doesn't actually teach the model how to use the existing knowledge more accurately.

We observe that although LLMs can memorize a vast range of world knowledge easily, they can also be attacked by counterfactual leading prompts since it's hard to learn how to use these world knowledge accurately[3]. Figure 1 shows an example. The LLM knows what the longest and second-longest rivers in the United States are. However, it hallucinates with a counterfactual leading prompt. In this paper, we introduce a counterfactual attack framework called HALLUATTACK which generates counterfactual instruction data to mitigate hallucinations. The basic idea is to induce LLMs to hallucinate on the knowledge they have already acquired. Firstly, given a LLM, we use factual prompts to collect its responses. These responses are guaranteed to be factually correct, which can indicate that this LLM has already learnt these knowledge from its training data. Then, given a factual response from the LLM, we use GPT-4 to generate counterfactual questions, which contain facts that conflict with this factual response from the LLM. After that, these counterfactual questions are used to attack the LLM. Those prompts which can make the LLM hallucinate will be used to generate instruction data. We use GPT-4 to generate the outputs of counterfactual prompts given encyclopedia documents as external evidence to guarantee both factuality and knowledge boundary of the outputs. Finally, we validate the instruction data generated by our HALLUATTACK by fine-tuning the attacked LLM. Compared with existing approaches, our approach is lightweighted with only simple fine-tuning, but can still improve the intrinsic factuality of the LLMs.

The contributions of this work are threefold:

- We propose a simple yet effective approach called HALLUATTACK to attack LLMs and generate counterfactual prompts which could make these LLMs hallucinate.
- We generate counterfactual instruction data by leveraging GPT-4 with encyclopedia documents as additional evidence. This instruction data can be further used to fine-tune the LLMs for hallucination mitigation.
- Experimental results on multiple open-source LLMs demonstrate the effectiveness and generalizability of our approach. The improvements on general LLM benchmarks also show the potential of counterfactual prompts on unlocking the LLM's self-reflection ability and better application of acquired world knowledge.

2 Related Work

2.1 Hallucination Detection and Mitigation

While the advancements in large language models(LLMs) have significantly elevated their performance across an array of downstream tasks, the issue of hallucination has emerged as a significant challenge. Hallucination is characterized by the generation of text by LLMs that deviates from the source material or fails to align with factual truthful information. These original texts and factual datasets typically serve as critical components in the training process, or as user-supplied prompts engaging with the LLMs.

(Huang et al., 2023a) proposes that hallucinations principally arise from three areas: the data source, the training phase, and the inferring phase. As a result, to effectively diminish the occurrence of hallucinations in the text generated by LLMs, a multitude of research has ventured into devising strategies for detecting and mitigating these hallucination problems across the aforementioned three areas.

Due to the potential presence of false factual information and biases in the data consumed by LLMs(Navigli et al., 2023), such as outdated or conflicting knowledge, and discrepancies between

[3]The knowledge gap due to insufficient data is beyond the scope of this work.

user prompts and the parametric knowledge in LLMs, hallucinations may occur. In response to this issue, a knowledge editing method was proposed by (Yao et al., 2023), which involves modifying the parametric knowledge of LLMs through the introduction of a model plug-in which similar to an adapter. Additionally, efforts have been made to mitigate hallucinations in LLMs by introducing high-quality, unbiased data through retrieval enhancement technology by (Lewis et al., 2020), (Guu et al., 2020), (Shi et al., 2023b). By refocusing LLMs on this reliable knowledge data, rather than potentially biased parameter knowledge, the hallucination rate of LLMs can be reduced.

A well-planned training and alignment strategy can help reduce the generation of LLMs hallucinations. A simple and effective hallucination elimination method named ICD (Zhang et al., 2024), which subtracts the output distribution of the induced Weak LLMs with hallucination problems from the output distribution of the original LLMs in training phase, thereby eliminating hallucinations to a certain extent. (Lee et al., 2022) introduced a fact-enhanced training method that significantly mitigates hallucination problems caused by differing factual information. Furthermore, in the LLMs alignment phase, (Wei et al., 2023) introduces simple synthetic data in an additional fine-tuning stage to enhance the model's independence from user opinions, thereby reducing the generation rate of hallucinations.

In the reasoning phase of the model, various studies have been conducted to detect and eliminate hallucinations. (Li et al., 2023) proposes a polling-based query method called POPE to detect visual object hallucination. (Zhang et al., 2023) introduces a hallucination detection method that does not require the introduction of external knowledge. (Manakul et al., 2023) detects hallucination through an idea that if an LLM has knowledge for a concept, sampled responses are likely to be similar. (Chuang et al., 2024) proposed a decoding strategy to reduce the hallucination of LLMs by comparing the logarithmic difference between the back layer and the front layer projected to the vocabulary space to obtain the distribution of the next word. Additionally, (Shi et al., 2023a) introduced context-aware decoding(CAD), which modifies the output distribution by reducing the reliance on prior knowledge, thereby encouraging the attention to overview information.

2.2 Counterfactual Tasks

Counterfactual tasks in the field of artificial intelligence refer to tasks that involve generating, comprehending, evaluating, and more under counterfactual conditions or assumptions. Counterfactual tasks emphasize inferring potential outcomes and effects by altering certain premises or conditions based on existing facts, which is essential for enhancing the ability of comprehending and reasoning effectively. (Xu et al., 2023) proposed a false information detection framework based on counterfactual reasoning, which can effectively detect biases in data source. (Ou et al., 2022) proposed a counterfactual-based open-domain dialogue data augmentation architecture called CAPT. (Rao et al., 2021) introduced an attention mechanism based on counterfactual, and evaluated the method on various fine-grained image recognition tasks, all of which showed significant improvements.

Furthermore, as LLMs continue to advance, research on counterfactual tasks integrated with LLMs is gaining momentum. (Wu et al., 2024) proposed an evaluation framework based on counterfactual tasks variants to explore the capabilities and limitations of LLMs. (Jin et al., 2023) generates an LLMs evaluation benchmark using causal reasoning and counterfactual reasoning. However, there are still many areas where counterfactual research on LLMs is not sufficient, especially in the detection and elimination of hallucinations.

3 Approach

3.1 Overview

We now provide an overview of our approach to explain the whole process and how different components interact with each other. As shown in Figure 2, our approach comprises three components[4]:

- **Factual Response Generation**, which aims to collect the learnt knowledge of a LLM.
- **Counterfactual Prompt Generation**, it aims to collect counterfactual prompts which can make the LLM hallucinate based on the factual responses.
- **Counterfactual Instruction Generation**, which aims to generate instruction data given the counterfactual prompts for LLM fine-tuning.

[4]All the prompts we used can be found in https://github.com/oldstree/halluattack

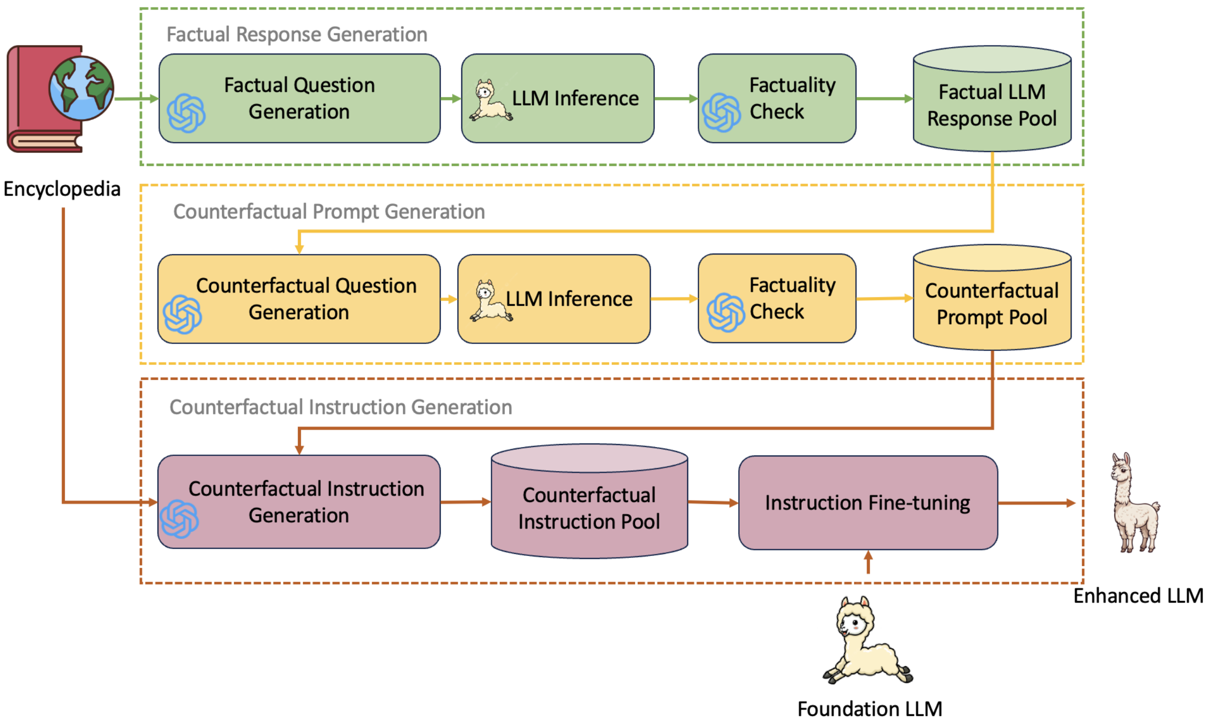

Figure 2: Overview of HALLUATTACK, comprising (1) Factual Response Generation, (2) Counterfactual Prompt Generation, and (3) Counterfactual Instruction Generation.

3.2 Factual Response Generation

There are many factors contributing to hallucinations in LLMs. As mentioned in Section 1, the hallucination factor of knowledge gaps due to insufficient data is beyond the scope of this work. So the first step of our approach is to know what the LLMs know. Based on this, we can then attack the LLMs, causing them to generate hallucination responses based on the knowledge they should have already mastered.

Firstly, we use GPT-4 to generate factual questions $FQ = \{fq_1, fq_2, ..., fq_k\}$ based on the provided encyclopedia document d_i (k factual questions for each encyclopedia document.). This step will guarantee that: a). The generated questions are knowledge-intensive, requiring the LLMs to answer using the knowledge they have learned. b). The generated questions come with background knowledge (the encyclopedia document) that can be used to help verify the correctness of the LLM's responses. c). When the LLMs answer incorrectly, the background knowledge can be utilized to generate factually correct responses.

Then, the LLM generates responses given the factual questions, and factuality check step using GPT-4 is applied to filter those factually correct responses $FR = \{fr_1, fr_2, ..., fr_m\}$. The encyclopedia text will be used as background knowledge for factuality check.

Example 1: Given an encyclopedia document of "List of rivers of the Americas[5]":

> *The Missouri River is the longest river in North America and the United States (2,341 mi (3,767 km)). The second longest river in North America and the United States is the Mississippi River (2,320 mi (3,730 km)).*

We will generate factual questions like *"What are the longest and second longest rivers in the United States?"*. One of the possible factual answer for this question is:

> *The longest river in the United States is the Missouri River, which is approximately ... The second longest river in the United States is the Mississippi River, which is approximately ...*

3.3 Counterfactual Prompt Generation

Given the factual responses, the main purpose of *Counterfactual Prompt Generation* is to find

[5] https://en.wikipedia.org/wiki/List_of_rivers_of_the_Americas

prompts which can attack the LLM and make it hallucinate. Similar as *Factual Response Generation*, given a factual response fr_j, and its corresponding encyclopedia text d_i, we use GPT-4 to generate counterfactual questions $CFQ = \{cfq_1, cfq_2, ..., cfq_k\}$ which contain conflict fact with the provided factual response. Then, these prompts will be used to attack the LLM. We check the factual correctness of the responses of these counterfactual questions by GPT-4 with the factual response fr_j and its corresponding encyclopedia text d_i as the background knowledge. Those prompts which can successfully make the LLM hallucinate will be left for next step.

Example 2: Given the factual response in *Example 1*, we can generate counterfactual questions like *"As the second longest river in the United States, which cities does the Missouri River flow through?"*

The LLM hallucinates on this question with responses like :

> *The Missouri River, which is the second longest river in the United States after the Mississippi River, flows through. . .*

3.4 Counterfactual Instruction Generation

Given a counterfactual prompt cfq_i, we need to generate high quality instruction data for further model fine-tuning. The instruction data should accurately identify the counterfactual errors in the prompts and should be as free of hallucinations as possible.

Instead of directly using super LLM's (e.g. GPT-4) responses as instruction data, given a counterfactual prompt, we incorporate its corresponding encyclopedia text d_i as the background knowledge to generate high quality responses using GPT-4. So we can minimize the hallucination of GPT-4 itself, thereby increasing the accuracy of the responses.

Example 3: Given the above counterfactual question in *Example 2*, the correct answer should be like:

> *Your question might be incorrect. The longest river in the United States is the Missouri River, which spans about 2,341 miles. The second longest river is the Mississippi River, which is approximately 2,320 miles long.*

3.5 Finetuning the LLM on the Counterfactual Instructions

Supervised Fine-tuing is a simple yet effective alignment method. Once the counterfactual instruction generation is done, we simply fine tune the attacked LLM with this data. We use the counterfactual prompts as the input to the LLM and require the model to generate the responses. A standard sequence-to-sequence loss is applied to train the LLM.

4 Experiments

4.1 Experimental Setup

In this section, we describe the data, models, and benchmarks of the experiments.

Corpora We use about 200,000 Chinese encyclopedia documents and generate 3,000 samples for each open source model for instruction tuning. The Chinese encyclopedia entries are sorted according to the popularity rank. Therefore, we can ensure that the encyclopedia documents used are definitely from the head portion and have certainly been utilized by the open-source LLMs.

Evaluation Models We evaluate our approach on several state-of-the-art LLMs, including Qwen1.5-7B-Chat[6], Qwen1.5-14B-Chat[7], Baichuan2-13B-Chat[8] and ChatGLM3-6B-32k[9].

Benchmark Datasets We select HalluQA[10](Cheng et al., 2023) and TruthfulQA(5-shot)[11](Lin et al., 2022) to evaluate the hallucination rate of the LLMs. We use the official evaluation scripts provided. Specifically, MC1 (Single-true) task is used for TruthfulQA.

In order to further evaluate the effectiveness of our approach on improving the LLM's ability of better using learnt knowledge, we also select several general LLM benchmarks including MMLU(Hendrycks et al., 2020), C-Eval(Huang et al., 2023b), GSM8K(Cobbe et al., 2021), BBH (Big Bench Hard)(Suzgun et al., 2022). We use OpenCompass[12] to evaluate the LLMs on these benchmarks which provides a comprehensive

[6] https://huggingface.co/Qwen/Qwen1.5-7B-Chat
[7] https://huggingface.co/Qwen/Qwen1.5-14B-Chat
[8] https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat
[9] https://huggingface.co/THUDM/chatglm3-6b-32k
[10] https://github.com/OpenMOSS/HalluQA/tree/main
[11] https://github.com/sylinrl/TruthfulQA/tree/main
[12] https://opencompass.org.cn/home

Table 1: Overall results on benchmarks of open-source model. *Imp.* denotes the improvement.

Model	C-Eval	MMLU	BBH	GSM8K	TruthfulQA	HalluQA
Qwen1.5-7B	68.88	61.50	40.35	55.57	53.85	42.88
+ HALLUATTACK	70.37	62.20	43.71	58.30	55.93	47.55
Imp.	**2.16%**	**1.14%**	**8.33%**	**4.91%**	**3.86%**	**10.89%**
Qwen1.5-14B	76.20	68.32	54.41	68.00	59.48	51.33
+ HALLUATTACK	76.90	68.45	56.46	70.43	60.34	52.22
Imp.	**0.92%**	**0.19%**	**3.77%**	**3.57%**	**1.45%**	**1.73%**
ChatGLM3-6B	52.12	50.79	41.25	24.11	35.98	31.33
+ HALLUATTACK	53.84	51.93	43.17	25.32	36.84	33.33
Imp.	**3.30%**	**2.24%**	**4.65%**	**5.02%**	**2.39%**	**6.38%**
Baichuan2-13B	56.31	59.17	48.78	52.77	45.65	45.77
+ HALLUATTACK	57.02	60.13	51.27	53.93	47.36	46.67
Imp.	**1.26%**	**1.62%**	**5.10%**	**2.20%**	**3.75%**	**1.97%**

benchmarking framework that enables us to systematically evaluate the performance of the LLMs across various tasks and domains.

Implementation Details We use GPT-4[13] as super LLM annotator in multiple components in our approach. We generate 3 facutual questions for each encyclopedia document and 3 counterfactual questions for each factual response.

We use Firefly[14], a open-source LLM fine-tuning framework for supervised fine-tuning of our evaluation models. Specifically, we employed a learning rate of 1e-5, a batch size of 4, and conducted training for ten epochs. Each model is trained on a single node with eight 80G NVIDIA A100 GPUs.

We utilize standard greedy decoding for inference to ensure the reproducibility. The maximum generation length is set to 1024.

4.2 Results

Table 1 presents a detailed comparison of various LLMs' performances on both general and hallucination benchmarks. Notably, our approach demonstrates a substantial improvement in reducing hallucinations on TruthfulQA and HalluQA. After fine-tuning with instruction data generated by our HALLUATTACK approach, the performance is significantly improved (with increases of up to 10%) compared with original chat models, which demonstrates the effectiveness of our approach in reducing the hallucinations of LLMs. Furthermore, we also observed gains on general LLM benchmarks, particularly on the BBH and GSM8K. This shows the potential of our counterfactual instruction tuning on unlocking the LLM's self-reflection ability and better application of acquired world knowledge.

Our approach achieved better performance on Qwen1.5-7B model compared with Qwen1.5-14B model. This phenomenon suggests that our approach is more effective on LLMs with smaller-scale. A plausible explanation is that LLMs with smaller-scale often struggle with robust reasoning capabilities and can hardly have a thorough understanding of knowledge boundaries. Our approach introduces the counterfactual instruction data. The data can detect where the knowledge boundaries of the LLMs are weak through counterfactual attack, and then repairs and enhances the knowledge boundaries in the alignment phase, which can strengthen the world knowledge learnt by LLMs and thus reduce hallucinations.

Furthermore, table 1 shows that our approach yielded much more significant enhancement on HalluQA as opposed to TruthfulQA across most LLMs. This is because our experimental corpus is derived from Chinese encyclopedic sources, offering a wealth of Chinese counterfactual data. Despite this, we still observed improvements on the English evaluate dataset (i.e. TruthfulQA). The phenomenon not only demonstrates the efficiency of our approach in leveraging linguistically and culturally specific datasets, but also shows the potential for hallucination reduction to be transferred across languages.

[13] https://platform.openai.com/docs/models/
[14] https://github.com/yangjianxin1/Firefly

4.3 Discussion

Corpora As mentioned before, we focus on improving the LLM's ability to better use the knowledge they've already acquired during pretraining phase. The knowledge gap due to insufficient data is beyond the scope of this work. So we deliberately use encyclopedia data which has already been used in pretraining phase to create counterfactual instruction data. No new knowledge will be introduced in supervised fine-tuning phase. Existing work(Wan et al., 2024) has shown that minimizing the inconsistency between external knowledge present in the alignment data and the intrinsic knowledge embedded within foundation LLMs is important for hallucination mitigation.

Instruction Generation As mentioned in section3.4, we use the original encyclopedia document as the background knowledge for GPT-4 to generate the output of the counterfactual prompt. This is very important to minimize the hallucination generated by GPT-4. However, this will probably change the generation behavior or style of the attacked LLM, because the output of the instruction data is mostly summarized from the given encyclopedia document, so the diversity and richness of the generated content will decrease. To tackle this challenge, we tried to use the factual response generated by LLM itself as another background knowledge. We hope the output of the counterfactual prompt can, on the one hand, point out the factual errors in the prompt, on the other hand, follow the original generation style as the factual response. However, the performance is not good as current setup in section3.4. After diving into several cases, we found that the quality generated by GPT-4 with two background documents is not very good, GPT-4 sometimes exhibits a mix and repetition of background documents, which may be due to the prompt we used. Moreover, there could be also some factual errors that are not easily detected automatically in the factual responses. If such data were used during the fine-tuning phase, it would actually exacerbate the LLM's hallucinations. How to improve the data quality and generate style consistent instruction data will be our future work to follow up.

Combination with other hallucination mitigation methods The proposed approach plays as a "patch" to given LLMs with simple continue fine-tuning. Since the data volume we used for fine-tuning is very small, we didn't observe catastrophic forgetting during fine-tuning. This implies that our approach can be integrated with existing hallucination mitigation approaches and can also serves as a supplement to them.

5 Conclusion

In this paper, we explore how counterfactual instruction data helps unlock the ability of LLMs to utilize knowledge more accurately, and propose a simple yet effective prompting approach to attack the LLMs and generate high quality counterfactual instruction data for model fine-tuning. Experimental results demonstrate the effectiveness and scalability of our approach in reducing hallucinations.

Limitations

In our approach, we leverage a super LLM, i.e. GPT-4, as annotators. Although the annotation tasks are not very complex (mostly are question generation and answer summarization tasks) and don't require huge world knowledge, it is still necessary to investigate more advanced approaches to improve the quality and diversity of the generation as mentioned in section4.3.

Our approach also achieved gains on general LLM benchmarks. We believe that the counterfactual instructions can unlock the self-reflection ability of the LLMs, thereby may improve the performance on any knowledge-intensive benchmarks. However, the underlying reasons have not yet been thoroughly explored. As a direction for future research, we propose to concentrate on the connections among counterfactual attack, Chain-of-thought(CoT)(Wei et al., 2022) and any other cognitive methods of LLMs. This should be essential for understanding the factuality and hallucination of LLMs.

Ethics Statement

All the data we used in the experiments are publicly available encyclopedia documents, which do not contain privacy information to the best of our knowledge.

We state that any research or application arising from this study is strictly authorized solely for research purposes.

References

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Conference on Empirical Methods in Natural Language Processing*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *CoRR*, abs/2310.03368.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *International conference on machine learning*. ICML.

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

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio C'esar Teodoro Mendes, Allison Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, S. Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuan-Fang Li. 2023. Textbooks are all you need. *ArXiv*, abs/2306.11644.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv*, abs/2311.05232.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 34586–34599. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *ArXiv*, abs/2401.03205.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 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)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rong-Zhi Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *ArXiv*, abs/2402.00253.

Junliang Luo, Tianyu Li, Di Wu, Michael R. M. Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation. *ArXiv*, abs/2401.08358.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding. *ArXiv*, abs/2305.04757.

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, Singapore. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Jiao Ou, Jinchao Zhang, Yang Feng, and Jie Zhou. 2022. Counterfactual data augmentation via perspective transition for open-domain dialogues. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1635–1648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1025–1034.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *ArXiv*, abs/2301.12652.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Conference on Empirical Methods in Natural Language Processing*.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zheng Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chun-Yan Li, Eric P. Xing, Furong Huang, Haodong Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Sekhar Jana, Tian-Xiang Chen, Tianming Liu, Tianying Zhou, William Wang, Xiang Li, Xiang-Yu Zhang, Xiao Wang, Xingyao Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models. *ArXiv*, abs/2401.05561.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Bo Tian, Siyuan Cheng, Xiaozhuan Liang, Ningyu Zhang, Yi Hu, Kouying Xue, Yanjie Gou, Xi Chen, and Huajun Chen. 2024. Instructedit: Instruction-based knowledge editing for large language models. *ArXiv*, abs/2402.16123.

S. M Towhidul Islam Tonmoy, S M Mehedi 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*, abs/2401.01313.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Fanqi Wan, Xinting Huang, Leyang Cui, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge verification to nip hallucination in the bud.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, Mexico City, Mexico. Association for Computational Linguistics.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789, Toronto, Canada. Association for Computational Linguistics.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Editing large language models: Problems, methods, and opportunities. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10222–10240, Singapore. Association for Computational Linguistics.

W. Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *ArXiv*, abs/2209.10063.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2024. Alleviating hallucinations of large language models through induced hallucinations.