The Dawn After the Dark: An Empirical Study on Factuality Hallucination in Large Language Models

Anonymous ACL submission

Abstract

In the era of large language models (LLMs), hallucination (*i.e.,* the tendency to generate factually incorrect content) poses great challenges to trustworthy and reliable deployment of LLMs in real-world applications. To tackle the hallucination, three key questions should be well studied: how to detect hallucinations (*detection*), why do LLMs hallucinate (*source*), and what can be done to mitigate them (*mitigation*). To address these challenges, this work presents a systematic empirical study on LLM hallucinations, focused on the three aspects of hallucination detection, source and mitigation. Specially, we construct a new hallucination benchmark *HaluEval 2.0*, and design a simple yet effective detection method for LLM hallucinations. Furthermore, we zoom into the different training or utilization stages of LLMs and extensively analyze the potential factors that lead to the LLM hallucinations. Finally, we implement and examine a series of widely used techniques to mitigate the hallucinations in LLMs. Our work has led to several important findings to understand the hallucination origin and mitigate the hallucinations in LLMs.

1 Introduction

Large language models (LLMs) (Zhao et al., 2023) have shown remarkable potential in a wide range of natural language processing applications (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023). However, despite the significant improvement in model capacity, a persistent challenge lies in their tendency to *hallucinate*, *i.e.,* generate content that looks plausible but is factually incorrect (Huang et al., 2023; Ji et al., 2023a; Zhang et al., 2023b). This issue severely restricts the deployment of LLMs in real-world applications (*e.g.,* clinical diagnoses), where the reliable generation of trustworthy text is of utmost importance.

In the era of LLMs, there has been a significant surge of research interests in hallucinations (Yao et al., 2023; Das et al., 2023; Dhuliawala et al., 2023a; Varshney et al., 2023; Manakul et al., 2023). These studies are mainly centered around three interleaved questions, *i.e., why do LLMs hallucinate, how to detect hallucinations, and what can be done to mitigate them?* The three key questions pose great challenges to the research community, while existing empirical work mostly focuses on analyzing or addressing individual challenges, still lacking a systematic and in-depth experimental study on LLM hallucinations. To bridge this gap, a more comprehensive analysis is needed to thoroughly research the aforementioned three questions.

For deciphering the mystery of hallucination in LLMs, we aim to conduct a comprehensive and systematic empirical study on hallucination *detection*, *source*, and *mitigation*. In particular, we mainly focus on studying *factuality hallucination*, which has become one of the primary erroneous sources for LLMs (Huang et al., 2023; Zhang et al., 2023b). To carry out our research, we zoom into the different stages to train and use LLMs, including pre-training, supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), and inference, and thus can conduct more in-depth analysis of potential impact of each stage on model hallucination. This analysis approach is quite different from prior work, where they mostly study the impact of individual stages or strategies to attribute or mitigate the hallucinations.

To conduct our empirical study, we first extend previous work (Li et al., 2023a) and construct a new benchmark **HaluEval 2.0** for evaluating the factuality hallucination in LLMs. Our benchmark contains 8,770 questions from five domains including biomedicine, finance, science, education, and open domain. To detect factual errors in LLM responses, we propose a simple yet effective framework that decomposes the task of hallucination detection into two simple sub-tasks, *i.e.,* extract factual statements from responses and judge the trustfulness of each

statement. Then, based on the detection approach, we perform a series of studies to explore the hallucination sources from four aspects, *i.e.,* pre-training, SFT, and inference methods. For each aspect, we extensively examine the effect of possible factors on LLM hallucinations. Finally, we delve into hallucination mitigation for LLMs through a series of widely used techniques, including RLHF, retrieval augmentation, and advanced decoding, etc.

It is generally considered to be challenging to deeply understand and fully mitigate LLM hallucinations. By providing more comprehensive empirical analysis, our work aims to further push forward the research on hallucinations. The major findings of our empirical study are summarized below:

• **Pre-training:** Pre-training on more tokens has an oscillatory effect on reducing LLM hallucinations. The familiarity of pre-training knowledge to LLMs significantly influences the source of hallucinations, *i.e.,* lower familiarity, more hallucinations.

• **Fine-tuning:** Supervised fine-tuning LLMs with improved and appropriate instructions can be useful to alleviate hallucinations. RLHF is an effective method to mitigate LLM halluciantions, while this effect relies on specific domains.

• **Inference:** Diversity-oriented decoding contributes to more hallucinations, which foregrounds the importance of balancing diversity and factuality during decoding. Augmenting generation with retrieval can effectively mitigate hallucinations.

2 Hallucination in LLMs

In the field of NLP, *hallucination* typically refers that the model output contains undesired content that is nonsensical or deviates from the source material (Ji et al., 2023a; Zhang et al., 2023b).

Despite that hallucination can be defined in different ways, we focus on *factuality hallucination*, since it has become one of the primary sources of erroneous responses by LLMs. In light of this research, we propose a fine-grained categorization of factuality hallucination in LLMs as follows:

• **Entity-error Hallucination.** This type of hallucination refers to the situations where the generated text of LLMs contains erroneous entities, such as person, date, location, and object, that contradict with the world knowledge.

• **Relation-error Hallucination.** This type of hallucination refers to the generated text of LLMs contains wrong relations between entities such as quantitative and chronological relation.

• **Incompleteness Hallucination.** LLMs might exhibit incomplete output when generating lengthy or listed responses. This hallucination arises when LLMs are asked about aggregated facts and they fail to reserve the factual completeness.

• **Outdatedness Hallucination.** This type of hallucination refers to situations where the generated content of LLMs is outdated for the present moment, but was correct at some point in the past. This issue arises primarily due to the fact that most LLMs were trained on time-limited corpora.

• **Overclaim Hallucination.** This type of hallucination means that the statement expressed in the generated text of LLMs is beyond the scale of factual knowledge (Schlichtkrull et al., 2023).

• **Unverifiability Hallucination.** In some cases, the information generated by LLMs cannot be verified by available information sources.

Note that it is difficult to encompass all kinds of hallucination, and our taxonomy aims to depict the most frequently occurring types of hallucination. We present several illustrative examples for each type of hallucinations in Table 7 in Appendix A.

3 Experimental Setup

In this section, we first introduce our benchmark and then describe a set of models for comparison.

3.1 Benchmark Construction

To comprehensively evaluate the tendency of LLMs to generate hallucinations across various domains, we extend the previous study of HaluEval (Li et al., 2023a) and meticulously construct an upgraded hallucination evaluation benchmark **HaluEval 2.0**. Our benchmark collects questions for five domains (*i.e.,* biomedicine, finance, science, education, and open domain) from six domain datasets, including BioASQ (Krithara et al., 2023), NFCorpus (Boteva et al., 2016), FiQA-2018 (Maia et al., 2018), SciFact (Wadden et al., 2020), LearningQ (Chen et al., 2018), and HotpotQA (Yang et al., 2018). The construction details of our benchmark are shown in Appendix B. In total, our benchmark consists of 8770 questions, with 1535, 1125, 1409, 1701, and 3000 questions for biomedicine, finance, science, education, and open domain, respectively.

3.2 Evaluation Models

We conduct the experiments with a number of representative open-source and closed-source LLMs based on HaluEval 2.0.

• *Open-source models.* We focus on instruction-tuned models, which use instructions (*e.g.,* daily chat, synthetic instructions) for tuning. In our experiments, we select six representative instruction-tuned models including Alpaca (7B) (Taori et al., 2023), Vicuna (7B/13B) (Chiang et al., 2023), and Llama 2-Chat (7B/13B) (Touvron et al., 2023a).

• *Closed-source models.* In contrast, closed-source models can be only accessed via APIs. Here, we select five exceptional closed-source models including text-davinci-002/003 (Ouyang et al., 2022), ChatGPT, Claude, and Claude 2.

4 Hallucination Detection

To analyze and mitigate hallucinations, the first and fundamental step is to detect hallucinations. In this section, we design an automatic hallucination detection approach and further validate its reliability by comparing with human labeling.

4.1 Detection Approach

We propose a simple yet effective framework to detect factual errors in model responses. Following prior work (Chern et al., 2023b; Dhuliawala et al., 2023b), we decompose the challenging hallucination detection task into two simple sub-tasks: 1) extract multiple factual statements from a lengthy response; and 2) determine whether each statement contains hallucinations.

Extraction-then-Verification. First, we instruct a LLM (*i.e.,* GPT-4) to extract factual statements that could be proven to be true or false based on the world knowledge. Instead of training a specific model for fact extraction (Thorne et al., 2018), this approach can significantly reduce the costs of data annotation and model training. The LLM might output "*NO FACTS*" if there is no factual statement in the response. Second, we use the LLM itself to judge the trustfulness of statements since the LLM has encoded rich world knowledge. Previous work checks each involved fact independently using separate prompts based on verification questions or external tools (Chern et al., 2023b; Dhuliawala et al., 2023b). However, we observe that these statements are often interrelated, with certain statements providing the background or serving as conditions for others. Thus, independently assessing each statement may lead to misjudgment of hallucination. In this work, we instruct the LLM with *all* statements to predict their hallucination judgements (*i.e., True*, *False*, or *Unknown*). To give a confident judgement, we only consider the *false* statement as hallucination in our following experiments. We present our used instructions in Appendix C.

Test of Reliability. To examine the reliability of our approach in hallucination detection, we invite human labelers to annotate the factuality of a subset from HaluEval 2.0 and compare the judgement of LLM and humans. Specially, based on the semantic similarity score (Section 3.1), we select 1,000 questions from HaluEval 2.0 with an average of 200 samples for each domain. For each question, we use ChatGPT to generate a response and ask human labelers to annotate the hallucination type of each statement (Section 2). To ensure the correctness of our annotation, each sample is labeled by two labelers and examined by one checker. Finally, the matching rate between the judgement of our proposed method and human annotation is 92.6%, 94.7%, 92.7%, 91.5%, and 93.9% for biomedicine, finance, science, education, and open domain, respectively. The high consistency demonstrates that our proposed method has a high level of reliability in detecting the hallucinations from LLMs.

Evaluation Metrics. In order to quantitatively evaluate LLMs on HaluEval 2.0, we design two metrics from different levels to measure the degree of LLMs generating hallucinations in their responses. The *micro hallucination rate (MiHR)* measures the proportion of hallucinatory statements within each response, which is calculated as:

$$\text{MiHR} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{Count}(hallucinatory\ facts)}{\text{Count}(all\ facts\ in\ r_i)}, \quad (260)$$

where n is the total number of samples in every domain and r_i is the i-th response. Besides, the *macro hallucination rate (MaHR)* calculates the proportion of responses containing hallucinatory statements, which is computed as:

$$\text{MaHR} = \frac{\text{Count}(hallucinatory\ responses)}{n}. \quad (266)$$

For both metrics, smaller value indicates better performance.

4.2 Results and Analysis

We evaluate LLMs on HaluEval 2.0 and apply our detection approach to measure their tendency to produce hallucinations in Table 1.

First, we can clearly observe that there exists a significant performance gap between open-source

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
ChatGPT	14.66	3.64	25.34	6.28	18.27	4.19	33.13	8.37	47.19	13.21
Claude 2	28.76	7.23	35.91	9.25	15.21	3.36	36.84	10.13	39.18*	12.62*
Claude	31.44	8.25	39.11	10.56	21.31	4.78	41.26	11.53	55.39	19.50
Text-Davinci-002	34.88	15.07	41.51	18.24	29.99	9.19	37.82	17.80	44.51	25.93
Text-Davinci-003	46.38	14.27	56.01	16.65	43.11	12.11	58.86	19.54	70.53	25.25
Vicuna 13B	50.59	17.55	46.19	13.15	34.44	8.75	55.81	17.88	65.43	29.15
Vicuna 7B	52.51	18.79	50.77	14.67	40.14	10.42	58.44	19.12	66.77	29.18
Llama 2-Chat 13B	52.61	17.90	53.48	14.53	39.11	10.37	62.12	19.30	79.19	30.44
Llama 2-Chat 7B	58.71	20.38	56.09	15.98	43.58	11.07	66.04	21.64	80.99	32.64
Alpaca 7B	53.52	24.42	53.47	24.46	40.74	12.74	68.95	22.38	65.65	29.57

Table 1: Evaluation results on the tendency of LLMs to generate hallucinations. "*" denotes that Claude 2 is overly cautious and excessively hedge innocuous requests, resulting in few valid responses and low hallucination rates.

and closed-source models. For open-source models, we can discover that scaling the model size can effectively decrease the tendency to generate hallucinations. Besides, we see that MaHR and MiHR are not strongly positively correlated (*e.g.,* Alpaca 7B vs Llama 2-Chat 7B), which might be due to the fact that some models tend to generate shorter responses with fewer facts. Furthermore, comparing the results across five domains, it shows that the tendency of LLMs to generate hallucinations is related to specific domains, *i.e.,* higher results in education and open domain. For open domain, we select the most difficult questions from HotpotQA where ChatGPT is likely to hallucinate. These findings suggest that the training methods of current LLMs in open domain are insufficient in preventing from generating hallucinations. Note that the percentage results of generating hallucinations in Table 1 might significantly exceed the actual rate in overall use, because our dataset is specially curated for hallucination evaluation.

Note: *The following experiments are based on the 1,000 samples selected in human annotation.*

5 Hallucination Source

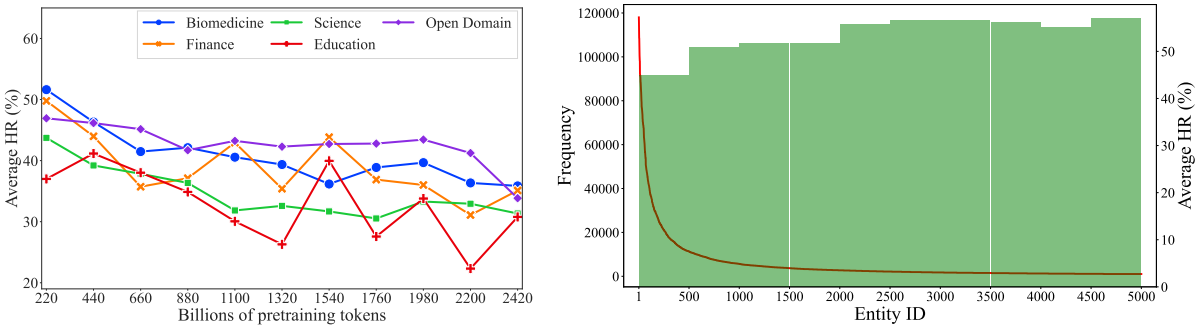
In this section, we perform a series of experiments to explore four factors that may induce LLM hallucinations, including pre-training, SFT, inference, and prompt design. We present some experiments of hallucination source in Appendix D.

5.1 Pre-training

Pre-training serves as the fundamental stage to establish the abilities of LLMs, enabling them to gain general capabilities and rich world knowledge. In this part, we consider two key factors, namely *scale* and *source* of pre-training data, since they are relatively easy to be examined according to public disclosure of LLMs and often have a large impact on model performance. Specially, we explore the effects of pre-training in three aspects related to the scale and source of pre-training data, *i.e.,* amount, familiarity, and mixture (see Appendix D.1).

Amount of Pre-training Tokens. We select 11 intermediate pre-training checkpoints of Baichuan 2 (7B) (Yang et al., 2023), corresponding to training on approximately 0.2 to 2.4 trillion tokens, and evaluate them on HaluEval 2.0. As shown in Figure 1 (a), with training on an increasing amount of tokens, the hallucination rate oscillates across these model checkpoints, suffering more in the domains of finance and education while overall decreasing in other three domains. This finding indicates that simply increasing the amount of pre-training tokens may not be that effective in hallucination reduction, which may require specific data strategies to alleviate the hallucinations in some specific domains.

Familiarity of Pre-training Knowledge. Prior work has reported that LLMs tend to produce hallucinations for those infrequent knowledge facts in pre-training corpus (Li et al., 2022). To explore this, we collect $5,000$ entities with the highest occurrence frequencies in Wikipedia (Karpukhin et al., 2020) and group them into 10 groups with descending frequencies. Finally, we randomly choose 100 entities for each group and construct a set of entity-based queries using manually designed templates. We evaluate Llama 2-Chat (7B) on these questions. As shown in Figure 1 (b), we can clearly observe a long-tail distribution on the entity frequency. Interestingly, the hallucination rates exhibit a clear three-level stair-like pattern on the 10 entity groups.



(a) Amount of Pre-training Tokens

(b) Familiarity of Pre-training Knowledge

Figure 1: (a) Average hallucination rate (%) of Baichuan 2 (7B) in five domains with respect to billions of pretraining tokens. (b) The red line denotes the frequency of entities, and the green bar denotes the average hallucination rate (%) of Llama 2-Chat (7B) for each group of entities.

Datasets	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
① FLAN-T5	73.12	32.33	70.29	28.79	45.00	18.33	67.55	32.73	64.44	31.48
② ShareGPT	66.11	25.34	68.25	23.92	42.21	11.66	66.67	25.00	62.20	28.33
③ Self-Instruct-52K	71.11	31.33	69.65	30.82	43.52	14.69	67.00	30.88	63.62	30.56
② + ③	70.17	29.60	67.22	24.02	39.69	12.52	66.94	27.76	56.56	29.43
① + ② + ③	64.52	26.80	70.00	25.29	43.81	13.56	67.14	28.63	56.67	31.56
③ Self-Instruct-52K	71.11	31.33	69.65	30.82	43.52	14.69	67.00	30.88	63.62	30.56
w/ complexity	69.74	31.08	62.70	21.17	42.21	13.02	69.34	30.21	55.40	29.34
w/ diversity	67.71	27.45	68.45	26.22	41.21	12.36	64.00	26.31	63.10	30.40
w/ scaling	65.98	27.80	67.63	29.24	37.50	12.42	64.10	29.59	53.33	29.96
w/ difficulty	57.38	25.78	65.87	30.86	31.47	12.20	49.62	27.89	42.96	24.06

Table 2: Evaluation results of LLaMA (7B) after SFT with different instruction datasets and instruction synthesis.

The LLM shows the lowest tendency to generate hallucinations for the most frequent entities in the first group. For the most entities in the long tail, the model exhibits a relative high hallucination rate.

5.2 Supervised Fine-Tuning

Supervised fine-tuning involves training language models to follow natural language instructions provided by the user. In this part, we explore the effect of instructions by mixing different types of instruction datasets and synthesizing new instructions.

Mixture of Instruction Datasets. Generally, there are three kinds of instructions, *i.e.,* task-specific, daily chat, and synthetic instructions. Hence, we select three representative instruction datasets, including FLAN-T5 (Chung et al., 2022), ShareGPT (Eccleston, 2023), and Self-Instruct-52K (Wang et al., 2022a), and then fine-tune LLaMA (7B) on each individual instruction set and their mixture to examine combinatorial effects. For a fair comparison we randomly sample 40K instructions for each instruction dataset. As shown in Table 2, daily chat instructions result in a lower level of hallucinations, while task-specific instructions lead to more hallucinations in responses. The task-specific instructions mainly focus on task format learning and ignore the factual knowledge. Mixing the daily chat and synthetic instructions can reduce hallucinations in some domains such as finance.

Types of Instruction Synthesis. To improve the capacities of LLMs, existing studies typically design a series of strategies to automatically construct large-scale instruction tuning data. Following prior work (Zhao et al., 2023), we consider four instruction synthesis methods based on Self-Instruct-52K: (1) *Enhancing the instruction complexity:* we adopt 40K instructions from WizardLM (Xu et al., 2023); (2) *Increasing the topic diversity:* we use ChatGPT to write 40K instructions for adapting to 293 topics; (3) *Scaling the instruction number:* we mix instructions from Moss (Sun et al., 2023) and Self-Instruct-52K to obtain 100K instructions; and (4) *Balancing the instruction difficulty:* we compute the perplexity of instructions by LLaMA (7B) to estimate the difficulty and remove too easy or too hard and keep 40K instructions. As shown in Table 2, fine-tuning

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
ChatGPT (greedy)	48.75	14.03	46.84	13.55	24.14	6.39	53.44	17.19	59.77	17.93
ChatGPT (top-p)	49.50	14.56	44.22	13.16	24.60	6.98	47.19	15.57	59.90	19.11
Llama 2 (greedy)	69.12	26.69	69.41	24.59	49.25	14.05	71.52	27.74	77.35	33.15
Llama 2 (top-p)	79.70	34.48	63.36	21.25	50.20	15.13	61.45	25.47	74.86	30.70
Llama 2 (top-k)	76.33	34.04	59.43	19.81	50.23	15.78	63.99	27.16	72.26	31.00
Llama 2 (beam)	73.13	32.47	59.54	21.06	47.00	14.55	63.35	25.24	73.84	29.22

Table 3: Evaluation results of different decoding strategies for ChatGPT and Llama 2-Chat (7B).

LLMs with improved instructions can be useful to alleviate hallucinations. Balancing the difficulty of instructions can significantly reduce hallucinations, while overly complex instructions eventually result in a higher level of hallucinations.

5.3 Inference Methods

In the inference stage, special decoding methods can be used to enhance the generation diversity but likely contribute to inducing hallucinations (Dziri et al., 2021). In this part, we will study the effects of decoding methods on hallucinations. Besides, the token-by-token generation manner (Zhang et al., 2023a) and quantization (Zhao et al., 2023) are also possible factors of inducing hallucinations, which are examined in Appendix D.2.

Decoding Strategies. We explore the influence of four decoding strategies, including greedy search, top-k sampling, top-p sampling, and beam search. Specifically, we test Llama 2-Chat (7B) and ChatGPT, and set $k = 20$, $p = 0.5$, and the number of beams to 5. For ChatGPT, due to the API constraint, we only investigate greedy search and top-p sampling strategies. The results are shown in Table 3. We observe that diversity-oriented decoding strategies such as top-p sampling contribute to inducing more hallucinations in professional domains (*e.g.,* science), while greedy search exacerbates hallucinations in open-ended domains (*e.g.,* education). This phenomenon is more pronounced in smaller models. Beam search can effectively balance the trade-off between the diversity and factuality.

6 Hallucination Mitigation

To alleviate the hallucinating behaviors of LLMs, several hallucination mitigation techniques have been proposed. In this part, we will study the effectiveness of three methods, including RLHF, retrieval augmentation, and advanced decoding. We also examine the effectiveness of self-reflexion and prompt improvement in Appendix E.

6.1 RLHF

RLHF is the process of fine-tuning language models with human feedback data to align with human values (Ouyang et al., 2022). Typically, employing RLHF to mitigate hallucinations involves two steps: (1) collect hallucinated and non-hallucinated responses to train a reward model; (2) fine-tune the LLM with the reward model using RL algorithms.

Experimental Details. We include the input questions in HaluEval 2.0 (besides the selected 1,000 test samples) as the initial prompt set. Following existing work (OpenAI, 2023), we pass a prompt to an unaligned LLM (*e.g.,* Alpaca 7B) and get a response, and feed the prompt and response through GPT-4 to rectify the hallucinations in the response (if any) with up to 5 rounds. We keep *(prompt, hallucinated response, refined response)* as comparison data to train a reward model. Finally, we fine-tune the unaligned LLM with the reward model using PPO. Specifically, we apply this approach to two unaligned LLMs, *i.e.,* Alpaca (7B) and Vicuna (7B), and evaluate aligned models on test samples.

Results and Analysis. As can be seen in Table 4, RLHF can effectively mitigate LLM hallucinations. Through the RLHF process, the model generates more accurate facts and its responses become more concise without much irrelevant content, leading to less factuality hallucinations. Moreover, the effect of RLHF is dependent on the domains, exhibiting more pronounced hallucination reduction in biomedicine and open domain, while showing mild effectiveness in highly professional domains such as science. In existing work, RLHF is mostly focused on open-ended domains but ignoring specific domains. We suggest that to make LLMs versatile it is crucial to execute RLHF in broader domains.

6.2 Retrieval Augmentation

Retrieval is generally considered as one of the effective approaches to alleviate hallucination (Li et al.,

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
Alpaca 7B	75.56	30.92	73.40	29.01	47.95	13.15	78.84	28.86	65.34	29.03
w/ RLHF	67.02	28.32	70.06	28.65	47.00	13.83	63.95	26.33	55.29	25.02
Vicuna 7B	72.59	27.75	73.06	25.28	49.49	13.79	70.78	27.62	64.52	30.31
w/ RLHF	70.85	25.90	70.33	23.87	48.50	13.45	68.32	25.39	53.75	25.78

Table 4: Evaluation results for Alpaca (7B) and Vicuna (7B) after RLHF alignment.

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
ChatGPT	48.75	14.03	46.84	13.55	24.14	6.39	53.44	17.19	59.77	17.93
w/ Retrieval	23.98	12.18	38.85	15.25	21.15	6.19	35.97	15.52	34.15	16.51
Llama 2-Chat 7B	69.12	26.69	69.41	24.59	49.25	14.05	71.52	27.74	77.35	33.15
w/ Retrieval	45.13	14.67	63.92	21.25	34.81	10.02	62.84	25.03	55.81	24.41
Llama 2-Chat 13B	70.56	26.63	69.95	23.85	42.21	13.02	69.14	26.71	76.34	32.48
w/ Retrieval	43.62	14.00	64.74	21.19	32.65	9.76	62.59	23.73	47.62	23.06

Table 5: Evaluation results for ChatGPT and Llama 2-Chat (7B and 13B) with retrieval augmentation.

2023a). The basic idea of retrieval augmentation is to first retrieve a small set of documents from a large-scale corpus (*e.g.,* Wikipedia) based on a user query and then the LLM can generate an accurate answer based on the retrieved documents.

Experimental Details. In our experiments, we conduct web search by retrieving documents from the whole web. Specially, we use the input question verbatim as query and request a call to Bing Search. For the search results, we only use the snippets of webpages as retrieved documents. Considering the context length and noise, we adopt top-2 snippets as evidence for generation. Appendix E.1 presents experiments about the impact of the number and relevance of retrieved documents on hallucinations.

Results and Analysis. As shown in Table 5, retrieval can significantly mitigate the hallucinations in LLM responses. The effectiveness of retrieval is more pronounced in smaller models (*e.g.,* Llama 2-Chat 7B), as the hallucination rate of larger models (*e.g.,* ChatGPT) has already been relatively low and smaller models acquire limited world knowledge. Figure 6 in Appendix E.1 shows that the lower the relevance between the retrieved document and question, the more likely the model is to produce hallucinations. However, more capable models like ChatGPT are more robust and less sensitive to the relevance of retrieved documents.

6.3 Advanced Decoding

We design two simple yet effective decoding methods that can flexibly switch between greedy search and top-p sampling to balance diversity and factuality. The first *greedy-nucleus sampling* assumes that when the model has high confidence in predicting the next word, it should adopt greedy search, otherwise use top-p sampling. The second *factual-nucleus sampling* (Lee et al., 2022a) hypothesizes that the randomness of generation will decrease as the model generating more words.

Experimental Details. We detail the two advanced decoding methods as follows:

• *Greedy-nucleus sampling*: Following previous studies (Li et al., 2023b), we use *entropy* to quantify the confidence of the model and set a confidence threshold η. We formulate this method as:

$$w_t = \begin{cases} \arg\max P(w_t|w_{1,\ldots,w_{t-1}}), & e \leq \eta \\ \sum P(w_t|w_{1,\ldots,w_{t-1}}) \geq p, & e > \eta \end{cases} \quad (1)$$

• *Factual-nucleus sampling*: Following previous work (Lee et al., 2022a), the probability p_t in top-p sampling at the t-th step can be formulated as:

$$p_t = \max\{\beta,\ p \times \lambda^{t-1}\}, \quad (2)$$

where λ is the decay factor that is used to decay the probability p at each step to reduce the randomness through time, and β is set as a lower bound of p to guarantee a certain degree of diversity. We reset p to the default value at the beginning of generating a new sentence. We apply the two advanced decoding methods to Llama 2-Chat (7B).

Results and Analysis. We present the results of the two advanced decoding methods in Table 6. As

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
greedy search	69.12	26.69	69.41	24.59	49.25	14.05	71.52	27.74	77.35	33.15
top-p sampling	79.70	34.48	63.36	21.25	49.20	13.13	61.45	25.47	74.86	30.70
greedy-nucleus	69.54	25.75	72.63	25.45	47.50	14.25	73.26	28.76	83.05	36.16
factual-nucleus	75.00	29.14	70.68	24.52	41.00	11.25	68.39	26.55	76.67	32.17

Table 6: Evaluation results of Llama 2-Chat (7B) using our designed greedy-nucleus and factual-nucleus sampling.

can be observed that, our proposed greedy-nucleus sampling strategy achieves comparable or even better results than the original decoding strategies in biomedicine and science, while the factual-nucleus sampling performs well in science, education, and open domains. It can be concluded that balancing the generation diversity and factuality benefits the reduction of hallucinations and the retention of text quality. However, our designed sampling methods are sensitive to hyper-parameters such as the confidence threshold and decay factor. Thus, we leave designing more effective decoding approaches to mitigating hallucinations in future work.

7 Related Work

Hallucination has been a fundamental challenge in LLMs, receiving extensive attention in existing literature (Huang et al., 2023; Ji et al., 2023a; Zhang et al., 2023b; Li et al., 2023a). We discuss them in two aspects: source/detection and mitigation.

Hallucination Source and Detection. To understand and detect the hallucination in LLMs, several studies focus on using the LLM itself as the tool to study the hallucinated content. For LLMs with access to their internal states, we can delve into the inner workings of the model to explore the principles behind hallucinations (Varshney et al., 2023; Yuksekgonul et al., 2023; Azaria and Mitchell, 2023a). Typically, the internal states that can be studied by examining the output logit values, the hidden layer activations, and the attention states. For example, Varshney et al. (2023) leveraged the output logit values of the model as a signal of hallucinations to estimate the uncertainty of responses. For models that can only be accessed through API calls, hallucinations are typically studied by analyzing the relationship between input prompts and the model's output responses (Rawte et al., 2023b; Manakul et al., 2023; Yao et al., 2023). For example, Rawte et al. (2023b) explored the effects of linguistic elements in prompts (readability, formality, and concreteness) on the LLM hallucinations.

Hallucination Mitigation. To mitigate the hallucinations, existing studies focus on different training and utilization stages of LLMs. First, mitigation in pre-training is typically centered around dataset curating and cleaning. In this line, existing studies (Das et al., 2023; Kamalloo et al., 2023; Umapathi et al., 2023) aim to construct a higher quality corpus for model pre-training by building datasets within specific domains or cleaning existing datasets. After pre-training, fine-tuning approaches can be further employed for hallucination mitigation, such as the applications of SFT (Wang et al., 2022b) and RLHF (Fernandes et al., 2023). In practical use, mitigation during generation is mainly focused on developing more effective decoding strategies (Lee et al., 2022b; Shi et al., 2023; van der Poel et al., 2022), leveraging external knowledge (Chern et al., 2023a; Varshney et al., 2023) and designing more effective prompts (Agrawal et al., 2023; Touvron et al., 2023b). Furthermore, mitigation during the post-processing stage can be implemented by using LLM itself (Mündler et al., 2023) or external knowledge (Chen et al., 2023) as the fact-checking module to verify the generated text.

8 Conclusion

This paper presented a comprehensive empirical analysis about LLM hallucinations in the three aspects of detection, source and mitigation. We constructed the hallucination benchmark HaluEval 2.0 and developed an LLM-based automatic detection approach. Based on this benchmark, we further systematically investigated the possible sources for LLM hallucination in the stages of pre-training, SFT, RLHF, and inference, and also examined the effectiveness of a series of hallucination mitigation strategies, including RLHF, retrieval augmentation, self-reflexion, advanced decoding, and prompt improvement. As the major contribution, our benchmark can be reused for further research, and our work has led to a series of important empirical findings on the source and mitigation of hallucination.

9 Limitations

Despite the great efforts that we have made, our analysis about pre-training stage and SFT is still limited, due to the lack of disclosed training details and the supporting computational sources. We will investigate into the two stages with more detailed analysis as future work. Furthermore, our experimental tests are not yet sufficient and we are also particularly interested in the working mechanism or nature of LLMs in generating the hallucinations. We will conduct more in-depth research work in the future. In addition, this paper mainly aims to provide empirical analysis on existing techniques to mitigate the LLM hallucinations, and there are no new hallucination mitigation strategies proposed. We will also consider developing improved mitigation strategies based on the findings of this work. Our hallucination detection approach is based on the GPT model (*i.e.,* GPT-4), which might inevitably lead to some minor errors. We will continuously improve our method in the future work.

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Amos Azaria and Tom Mitchell. 2023a. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Amos Azaria and Tom M. Mitchell. 2023b. The internal state of an LLM knows when its lying. *CoRR*, abs/2304.13734.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *CoRR*, abs/2204.06745.

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*, pages 716–722. Springer.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Guanliang Chen, Jie Yang, Claudia Hauff, and Geert-Jan Houben. 2018. Learningq: A large-scale dataset for educational question generation. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*, pages 481–490. AAAI Press.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023a. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023b. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei.

2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Souvik Das, Sougata Saha, and Rohini K Srihari. 2023. Diving deep into modes of fact hallucinations in dialogue systems. *arXiv preprint arXiv:2301.04449*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023a. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023b. Chain-of-verification reduces hallucination in large language models. *CoRR*, abs/2309.11495.

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2197–2214. Association for Computational Linguistics.

Dom Eccleston. 2023. Sharegpt. https://sharegpt.com/.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*.

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

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating hallucination in large language models via self-reflection. *CoRR*, abs/2310.06271.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv preprint arXiv:2307.16883*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10:170.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022a. Factuality enhanced language models for open-ended text generation. In *NeurIPS*.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *CoRR*, abs/2305.11747.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jingyuan Wang, Jian-Yun Nie, and Ji-Rong Wen. 2023b. The web can be your oyster for improving language models. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 728–746. Association for Computational Linguistics.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022. How pre-trained language models capture factual knowledge? A causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1720–1732. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

OpenAI. 2023. Gpt-4 technical report. *OpenAI*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Vipula Rawte, Prachi Priya, S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Amit P. Sheth, and Amitava Das. 2023a. Exploring the relationship between LLM hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *CoRR*, abs/2309.11064.

Vipula Rawte, Prachi Priya, SM Tonmoy, SM Zaman, Amit Sheth, and Amitava Das. 2023b. Exploring the relationship between llm hallucinations and prompt linguistic nuances: Readability, formality, and concreteness. *arXiv preprint arXiv:2309.11064*.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. Teler: A general taxonomy of LLM prompts for benchmarking complex tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14197–14203. Association for Computational Linguistics.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8618–8642. Association for Computational Linguistics.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023. Moss: Training conversational language models from synthetic data.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. *arXiv preprint arXiv:2210.13210*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *CoRR*, abs/2308.13259.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language model with self generated instructions. *CoRR*, abs/2212.10560.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022b. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *CoRR*, abs/2304.12244.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, and Li Yuan. 2023. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Mert Yuksekgonul, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. 2023. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. *arXiv preprint arXiv:2309.15098*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball. *CoRR*, abs/2305.13534.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Appendix

A Hallucination Taxonomy

In Section 2, we have defined six types of factuality hallucination in a fine-grained way. To make a clear understanding of our categorization of LLM hallucinations, we present some examples for each type in Table 7.

B Benchmark Construction

Specially, we firstly extract fact-intensive questions from six widely-used domain datasets, including BioASQ (Krithara et al., 2023), NFCorpus (Boteva et al., 2016), FiQA-2018 (Maia et al., 2018), SciFact (Wadden et al., 2020), LearningQ (Chen et al., 2018), and HotpotQA (Yang et al., 2018). To attain a sufficient number of high-quality questions, we adopt the training set of BioASQ and FiQA-2018, the whole set of NFCorpus and SciFact, and the test set of LearningQ (TED-Ed) and HotpotQA. Then, we only keep input questions ended with the question mark, and for SciFact we transform the input statements into questions using manually designed templates (*e.g., [statement] Can you provide some explanations to this statement?*). Finally, following HaluEval (Li et al., 2023a) to select questions that LLMs are most likely to hallucinate, we employ ChatGPT to generate three responses for each question and compute their average semantic similarity via BERTScore (Zhang et al., 2020). We only retain those questions where the similarity score is lower than a threshold. In our benchmark, the ques-

Type	Question	LLM Response
Entity-error Hallucination	When was the Nokia 3510 phone announced?	The Nokia 3510 mobile phone was announced in October 2002. It was a popular mobile phone atthe time, known for its compact design and features such as a color display and customizablecovers.
	Which merican voice actor wrote Iron Man: Rise of Technovore?	The American voice actor involved in English dubs of Japanese anime as well as cartoons who wrote "Iron Man: Rise of Technovore" is Brendan Burch.
Relation-error Hallucination	What are the chemical elements that make up the human body?	There are dozens of chemical elements that make up the human body, including hydrogen, oxygen, calcium, iron, and potassium. Among these chemical elements, calcium is more than oxygen.
	Who was born first, Aaron Gillespie or Nathan Leone?	Aaron Gillespie was born before Nathan Leone.
Incompleteness Hallucination	List ten book titles on social cognitive theory	1."Social Cognitive Theory: An Agentic Perspective" by Albert Bandura. ... 8."Cognitive Foundations of Soclal Behavior" by John L, Hayes and Paul J. Gendolla
	What is the components of an ecosystem?	An ecosystem consists of non-living matter and energy, producers and consumers, where the producers are mainly plants and the consumers are mainly people, animals, etc.
Outdatedness Hallucination	Who is the present president of the United States?	The present president of the United States is Donald Trump.
	Which team wins the 2022 FIFA World Cup?	France won the match 4–2, claiming their second World Cup and becoming the fourth consecutive title won by a European team, after Italy in 2006, Spain in 2010, and Germany in 2014.
Overclaim Hallucination	What is the validity of the assertion that "TMEM27 is a marker for beta cells."?	A study suggested that TMEM27 is expressed in purified human beta cells but not in other non-beta pancreatic cells.
	How to lose weight?	The only way to lose weight is to exercise, so you need to schedule many fat loss exercises to achieve your weight loss goals.
Unverifiability Hallucination	Do you agree that ten percent of kidney cancers in the UK are attributable to excess weight?	The World Health Organization has stated that being overweight or obese increases the risk of developing kidney cancer.
	List ten book titles on social cognitive theory	1."Social Cognitive Theory: An Agentic Perspective" by Albert Bandura. ... 8."Cognitive Foundations of Soclal Behavior" by John L, Hayes and Paul J. Gendolla.

Table 7: Illustrative examples for six types of LLM hallucinations that we define. We use red color to denote the incorrect content that contradicts with the world knowledge and unverifiable content that cannot be proven to be true or false by existing available sources.

You are presented with an answer to a user's query. Your task is to extract up to five factual statements from the answer. The extracted factual statement can be proven to be true or false through objective evidence. Please generate non-repetitive factual statements that can be independently verified. If the answer does not contain any factual statements, your response should be "NO FACTS". Here are four examples:

Context: <query>: Tell me a short bio of Franklin D. Roosevelt. <answer>: Franklin Delano Roosevelt, commonly known as FDR, was born on January 30, 1882, in Hyde Park, New York. In 1905, he married his distant cousin, Eleanor Roosevelt. FDR embarked on a political career in 1911, starting as a New York state senator. He then served as Assistant Secretary of the Navy under President Woodrow Wilson during World War I.
Response:
1. Franklin Delano Roosevelt was born in Hyde Park, New York on January 30, 1882.
2. FDR married his distant cousin, Eleanor Roosevelt in 1905.
3. In 1911, Franklin Delano Roosevelt started his political career as a New York state senator.
4. Franklin Delano Roosevelt served as Assistant Secretary of the Navy under President Woodrow Wilson during World War I.

Context: <query>: What was the outcome of the Battle of Waterloo? <answer>: The Battle of Waterloo was a significant battle fought in 1815. It marked the final defeat of Napoleon Bonaparte. The victorious Allied forces, led by the Duke of Wellington and Marshal Blücher, defeated Napoleon's French army. This victory ultimately resulted in the exile of Napoleon to the island of Saint Helena.
Response:
1. The Battle of Waterloo took place in 1815.
2. The Battle of Waterloo marked the final defeat of Napoleon Bonaparte.
3. The Duke of Wellington and Marshal Blücher commanded the victorious Allied forces.
4. Napoleon Bonaparte was exiled to the island of Saint Helena after the Battle of Waterloo.

Context: <query>: What is the capital of France? <answer>: The capital of France is Paris.
Response:
1. Paris is the capital of France.

Context: <query>: Create a website title and description for a website that provides free online courses. <answer>: Title: Free Learning Hub: Your Gateway to Online Knowledge Description: Welcome to Free Learning Hub – your one-stop destination for free online courses on various topics. Our website offers a wide range of learning opportunities that can help you develop new skills, improve your knowledge, and advance your career. From IT and programming to business management and language learning, we have something for everyone. Our courses are designed by industry experts and are accessible 24/7 from anywhere in the world. Join us today and unlock your potential with the power of free online learning!
Response:
NO FACTS

Context: <query>: {query} <answer>: {answer}
Response:

Table 8: Instruction of fact extraction for hallucination detection. The blue text denotes the intention description, and the green text denotes the in-context demonstration.

tions in biomedicine and science are expert-written, specialized, and require at least high-school knowledge to answer, while the questions in other domains are open-ended, mainly collected from daily conversation and open websites. In open domain, the questions from HotpotQA are based on the factual knowledge from Wikipedia.

C Hallucination Detection

Our hallucination detection approach consists of two steps: fact extraction and fact verification. The instructions for the two steps are shown in Table 8 and Table 9.

D Hallucination Source

D.1 Pre-training

Mixture of Pre-training Corpus. Existing LLMs typically employ a mixture of diverse public textual datasets as the pre-training corpus. The distribution of data source will affect the acquisition of general and domain-specific knowledge. To examine the effect of pre-training data mixture on hallucinations, we evaluate three LLMs with similar model sizes but using different pre-training corpus of general data (*e.g.,* webpages) and specialized data (*e.g.,* scientific text): Falcon (40B) (Almazrouei et al., 2023), Galactica (30B) (Taylor et al., 2022), and GPT-NeoX (20B) (Black et al., 2022). Their pre-training data distribution is shown in Figure 2 and our results are shown in Table 10. We can see that training on scientific data can significantly prevent the model from generating hallucinations in science and open domain (Galactica v.s. GPT-NeoX). General webpages data benefits reducing hallucinations in the domains of finance, science, and education. On the other hand, training on diverse corpora (*i.e.,* GPT-NeoX) instead tends to result in much more

You are given a list of statements extracted from a passage. Your task is to determine whether each statement is true or false in order and provide corrections for any false statements. If some statements are vague and difficult to determine, answer "UNKNOWN". Here are three examples:

Context: <statements>:
1. Franklin Delano Roosevelt was born in Hyde Park, New York on January 30, 1882.
2. FDR married his distant cousin, Eleanor Roosevelt in 1905.
3. In 1911, Franklin Delano Roosevelt started his political career as a New York state senator.
4. Franklin Delano Roosevelt served as Assistant Secretary of the Navy under President Woodrow Wilson during World War I.
Response:
1. FALSE [correction]: Franklin Delano Roosevelt was born in Hyde Park, New York on January 30, 1882.
2. FALSE [correction]: Franklin Delano Roosevelt married his distant cousin, Eleanor Roosevelt in 1905.
3. UNKNOWN
4. TRUE

Context: <statement>:
1. The Battle of Waterloo took place in 1815.
2. The Battle of Waterloo marked the final defeat of Napoleon Bonaparte.
3. The Duke of Wellington and Marshal Blücher commanded the victorious Allied forces.
4. Napoleon Bonaparte was exiled to the island of Saint Helena after the Battle of Waterloo.
Response:
1. TRUE
2. TRUE
3. FALSE [correction]: Napoleon Bonaparte was exiled to the island of Saint Helena after the Battle of Waterloo.
4. UNKNOWN

Context: <statement>:
1. Paris is the capital of France.
Response:
1. TRUE

Context: <statements>: {facts}
Response:

Table 9: Instruction of fact verification for hallucination detection. The blue text denotes the intention description, and the green text denotes the in-context demonstration.

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
Falcon 40B	53.25	20.25	39.61	15.35	36.75	10.89	47.83	22.82	61.96	28.94
Galactica 30B	51.70	16.51	58.93	23.18	41.72	11.24	51.76	18.09	52.17	25.25
GPT-NeoX 20B	63.27	22.29	60.58	23.38	46.95	14.28	61.17	23.77	61.62	29.26

Table 10: Evaluation results in five domains for three pre-trained models with diverse pre-training corpus.

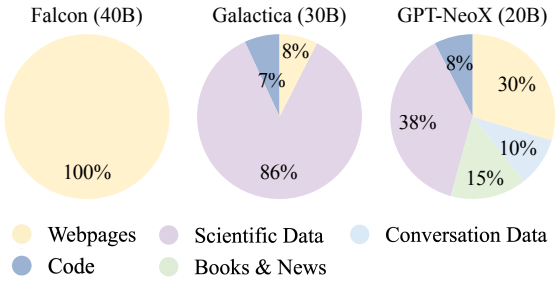
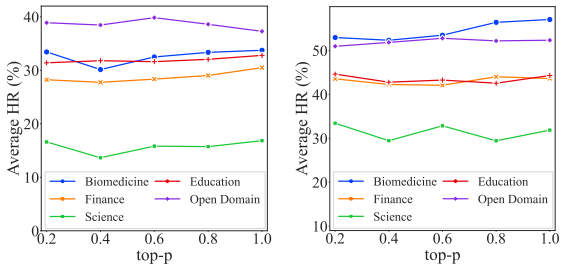


Figure 2: Ratios of pre-training data sources (figure copied from the LLM survey article (Zhao et al., 2023)).



(a) ChatGPT (b) Llama 2-Chat (7B)

Figure 3: Average hallucination rate (%) with varying p in top-p sampling.

hallucinations. Note that a potential factor to affect our results is that the data cleaning procedure for each model, since it would produce pre-training data of different data qualities. We leave this discussion of data cleaning in future work.

D.2 Inference Methods

Decoding Strategies. We vary p of top-p sampling in the set $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ and show

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
Llama 2-Chat 7B										
INT 16 (original)	69.12	26.69	69.41	24.59	49.25	14.05	71.52	27.74	77.35	33.15
INT 8	76.84	33.96	70.21	23.13	50.50	13.80	78.53	26.77	73.73	30.36
INT 4	76.16	31.80	67.54	23.08	48.50	14.30	79.97	30.00	79.41	35.00
Llama 2-Chat 13B										
INT 16 (original)	70.56	26.63	69.95	23.85	42.21	13.02	69.14	26.71	76.34	32.48
INT 8	75.66	32.43	69.53	23.77	45.20	14.38	69.94	26.74	78.57	33.70
INT 4	77.96	33.84	68.49	23.64	44.20	14.21	74.71	27.70	79.12	33.71

Table 11: Evaluation results for quantized Llama 2-Chat (7B and 13B).

Domain	Question	ChatGPT Response
Biomedicine	Is the regulation of Vsr endonuclease independent of the growth phase of bacteria?	During the exponential growth phase, when bacteria are actively dividing and replicating their DNA, DNA replication errors and mismatched bases can **more** frequently occur. These errors can lead to mutations in the genetic code, introducing variations in the bacterial population. ✔
		During the exponential growth phase, when bacteria are actively dividing and replicating their DNA, DNA replication errors and mismatched bases can **less** frequently lead to significant consequences. This is because the rapid pace of cell division during the exponential growth phase may outpace the usual repair mechanisms that cells employ to correct such errors. ✗
Finance	Record retention requirements for individuals in the U.S.?	Generally, it is suggested to keep tax-related records for **at least** seven years. This duration allows individuals and businesses to comply with various tax regulations and be prepared for any potential audits or inquiries from tax authorities. ✔
		Generally, it is suggested to keep tax-related records for **at most** seven years. This timeframe allows individuals and businesses to comply with the typical statute of limitations for tax audits and inquiries. ✗
Science	Do you agree with the claim that men are more susceptible to death due to pneumonia when compared to women?	The gender gap in pneumonia mortality becomes more prominent with age. In older age groups, **men** have a higher risk of dying from pneumonia compared to women. ✔
		The gender gap in pneumonia mortality becomes more prominent with age. In older age groups, **women** exhibit a higher susceptibility to severe pneumonia-related outcomes compared to men. ✗
Education	what innovation did the roman numerals use ?	The Roman numeral system introduced subtractive notation, where a **smaller** value is placed before a larger value to indicate subtraction. ✔
		The Roman numeral system introduced subtractive notation, where a **larger** value is placed before a smaller one to indicate subtraction. ✗
Open Domain	Can you give some details about Pluto?	Pluto is the **second smallest** dwarf planet in our solar system after Eris. ✔
		Pluto is the **smallest** and farthest known dwarf planet in our solar system, discovered by astronomer Clyde Tombaugh in 1930. ✗

Table 12: Illustrative examples for token-by-token generation showing that the LLM will commit to the previously generated token even if it might lead to hallucinations. The brown span denotes the same generated prefix by ChatGPT, and the **bold** font denotes the replaced token.

the trend in Figure 3. As can be seen, the hallucination rates of ChatGPT and Llama 2-Chat are highly sensitive to the variance of p within the professional domains, oscillating in biomedicine and science, while exhibiting minimal fluctuations in other open-ended domains.

Quantization. To understand the impact of quantization on hallucinations, we quantize Llama 2-Chat (7B) and (13B) at three precision levels: 4-bit, 8-bit and 16-bit. Specially, we employ the library

`bitesandbytes`[1] to quantize the original 16-bit models to 8/4 bits by specifying the commands `load_in_8bit` and `load_in_4bit`, which focused on the quantization of weights for LLMs. The hallucination results of quantized models are shown in Table 11. Despite with reduced memory footprint and accelerated inference rate, the 8-bit and 4-bit quantization results in an overall higher level of hallucination compared to the original 16-bit model. In most cases, 8-bit quantization has a minimal impact on the hallucination of the model, while 4-bit quantization significantly increases the hallucination in the model's responses. In certain domains such as biomedicine, the hallucination rate of the quantized model will noticeably increases compared to the original model.

Token-by-Token Generation. Most LLMs adopt the token-by-token manner to generate a token at a time. Once the generated part contains erroneous or unreasonable content, the model is difficult to complete the sentence correctly (Azaria and Mitchell, 2023b). To explore the drawback of token-by-token generation, we conduct a case study to qualitatively analyze how the generation paradigm leads to hallucinations. We present several examples across five domains in Table 12. For each sample, we employ ChatGPT to generate two responses conditioned on two similar prefixes with the only difference being the last word. We can clearly observe that the LLM over-commits the errors in previous tokens and continuously complete the sentence incorrectly. For example, when we replace the degree words (*e.g.,* "*more*" → "*less*"), ChatGPT does not identify the mistakes and keep generating the same sequence, leading to hallucinations. Besides, for the educational question about the roman numerals, when beginning with a prefix word "*larger*", ChatGPT is unable to indicate subtraction by generating the right word "*after*" (instead of "*before*"). Similarly, in the scientific domain, the real-world fact is that men have a higher risk of dying from pneumonia than woman. When we replace the prefix word "*men*" with "*woman*", ChatGPT is not able to complete the fact correctly by exchanging another expression. In open domain, when we delete "*second*" from the first prefix, ChatGPT incorrectly completes the sentence. In fact, one correct completion for the new prefix is "*Pluto is the smallest celestial body in the solar system that has ever been classified as a planet*".

[1]https://github.com/TimDettmers/bitsandbytes

D.3 Prompt Design

Prompting has become the major approach to utilizing LLMs. However, inappropriate prompt design would lead to incapable attention of important information in the input (Liu et al., 2023). In addition, ambiguous and superficial questions posed by users might steer the model towards generating unrelated, implausible, or bizarre output (Rawte et al., 2023a). In this part, we continue to analyze the effect of prompt design on LLM hallucinations.

Prompt Design. Generally, a prompt contains task description, input question, and contextual information such as in-context demonstrations (Santu and Feng, 2023). Here, we experiment with several prompt designs by varying the three ingredients:

• *Base prompt*: the initial prompt with a simple task description and input question.

• *Manual description prompt*: manually rewriting the task description in base prompt.

• *Synthetic description prompt*: using ChatGPT to synthesize the task description in base prompt.

• *Refined question prompt*: refining the initial question in base prompt by ChatGPT.

• *Manual in-context prompt*: manually selecting five in-context demonstrations for the base prompt.

• *Retrieved in-context prompt*: retrieving demonstrations based on BERT similarity from HaluEval 2.0 (besides the 1,000 test samples).

• *Synthetic in-context prompt*: using ChatGPT to synthesize the in-context demonstrations.

• *Reverse prompt*: reversing the position of task description and input question in base prompt (*i.e.,* place the task description after the input question).

We feed these prompts into ChatGPT and Llama 2-Chat (7B) and the evaluation results are shown in Table 13. First, we can observe that rewriting the task description with more details can reduce the hallucinations to some extent, while this effect is varied in domains. For professional domains (*i.e.,* biomedicine and science), incorporating more details into the task description can mitigate some hallucinations. Second, leveraging in-context learning can also help eliminate hallucinations in LLM's responses. These examplars or demonstrations can be manually selected, retrieved from candidate corpus, or automatically generated by LLM itself. It is noting that MaHR and MiHR are not strongly positively correlated, where the two metrics measure the hallucination degree from distinct levels. Finally, in most cases, rewriting the question or placing the task description at the end of the input

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
ChatGPT										
w/ base prompt	48.75	14.03	46.84	13.55	24.14	6.39	53.44	17.19	59.77	17.93
+ manual desc	45.64	13.91	39.20	11.18	22.34	5.28	55.68	17.73	64.52	20.31
+ synthetic desc	51.00	14.23	44.33	11.93	25.00	6.15	55.87	17.79	52.02	17.20
+ refined question	50.76	14.89	44.16	12.11	25.25	6.29	53.19	16.27	64.36	20.27
+ manual demo	42.71	14.89	40.74	12.12	25.27	7.11	56.41	19.24	44.72	21.88
+ retrieved demo	46.52	16.02	42.78	11.62	19.59	4.87	51.25	18.86	50.34	20.66
+ synthetic demo	38.10	18.30	36.69	15.13	27.71	8.15	43.90	23.24	29.17	18.08
+ reverse position	54.82	15.67	48.22	13.98	26.77	6.67	51.60	16.94	67.21	21.10
Llama 2-Chat 7B										
w/ base prompt	69.12	26.69	69.41	24.59	49.25	14.05	71.52	27.74	77.35	33.15
+ manual desc	68.02	26.46	74.36	25.01	42.50	12.10	76.16	30.97	79.39	33.23
+ synthetic desc	75.25	29.56	66.33	23.27	41.00	12.02	72.16	29.31	78.45	34.50
+ refined question	74.87	31.35	68.02	25.04	44.00	13.22	72.83	29.50	81.92	35.11
+ manual demo	69.70	27.90	66.33	24.61	45.00	12.27	71.01	27.02	66.88	31.84
+ retrieved demo	66.33	26.94	72.36	26.31	42.50	13.19	70.06	29.76	67.24	35.56
+ synthetic demo	59.68	27.31	62.84	24.43	45.50	14.72	57.64	27.49	53.77	30.27
+ reverse position	70.92	29.52	75.39	26.35	41.00	12.22	71.51	29.40	73.89	32.29

Table 13: Evaluation results of ChatGPT and Llama 2-Chat (7B) using different prompt formats.

question will instead hurt the model performance and induces more hallucinations.

Question Content. Following prior work (Rawte et al., 2023a), we further delve into how linguistics of question content, specifically readability, formality, and concreteness, influence the occurrence of LLM hallucinations. *Readability* quantifies the extent to which the question can be understood by humans; *Formality* refers to the degree of appropriate tone and professionalism conveyed by the choice of words, grammatical structure, style, etc.; *Concreteness* indicates whether a word represents a specific and tangible concept. For each question in our dataset, we invite human labelers to score these three properties based on the 5-point Likert scale, ranging from 1-point ("very terrible") to 5-point ("very satisfying"). For these questions, we instruct ChatGPT to generate responses and compute the average hallucination rate under each score. In Figure 4, we can observe that ChatGPT exhibits a lower propensity to generate hallucinations on those questions that are easier to read and use more formal and specific language. Note that constrained by the scale of the annotation dataset, there were fewer/no scores from humans for certain linguistics (*e.g.,* "2-point" formality in biomedicine), resulting in a relatively low hallucination rate.

E Hallucination Mitigation

E.1 Retrieval Augmentation

Experimental Details. To examine the impact of the number of retrieved documents on hallucination mitigation, we use the top-k documents as context and vary k in the set $\{1, 2, 5, 10\}$. Besides, to further validate the effect of the relevance of retrieved documents on hallucination mitigation, we randomly sample one document from top-k documents. Here, the variance of k reflects four levels of relevance to the question, ranging from strong to weak relevance.

Results and Analysis. In Figure 5, we show the results of using top-k documents as evidence. We can clearly see that ChatGPT and Llama 2-Chat mostly produce less hallucinations in top-2 retrieval, except that ChatGPT prefers top-5 in science. This is because that including more documents as the context of LLMs can bring more noises into the generation process, leading to higher level of hallucination. Therefore, we conduct top-2 retrieval for several LLMs and present the results in Table 5. Besides, we present the effect of the relevance between question and document by randomly selecting one document from top-k results. The results in Figure 6 shows that the lower the relevance between the retrieved document and the question, the more likely the model is to generate hallucinations. However, more capable models like ChatGPT are less sensitive to the relevance of retrieved docu-

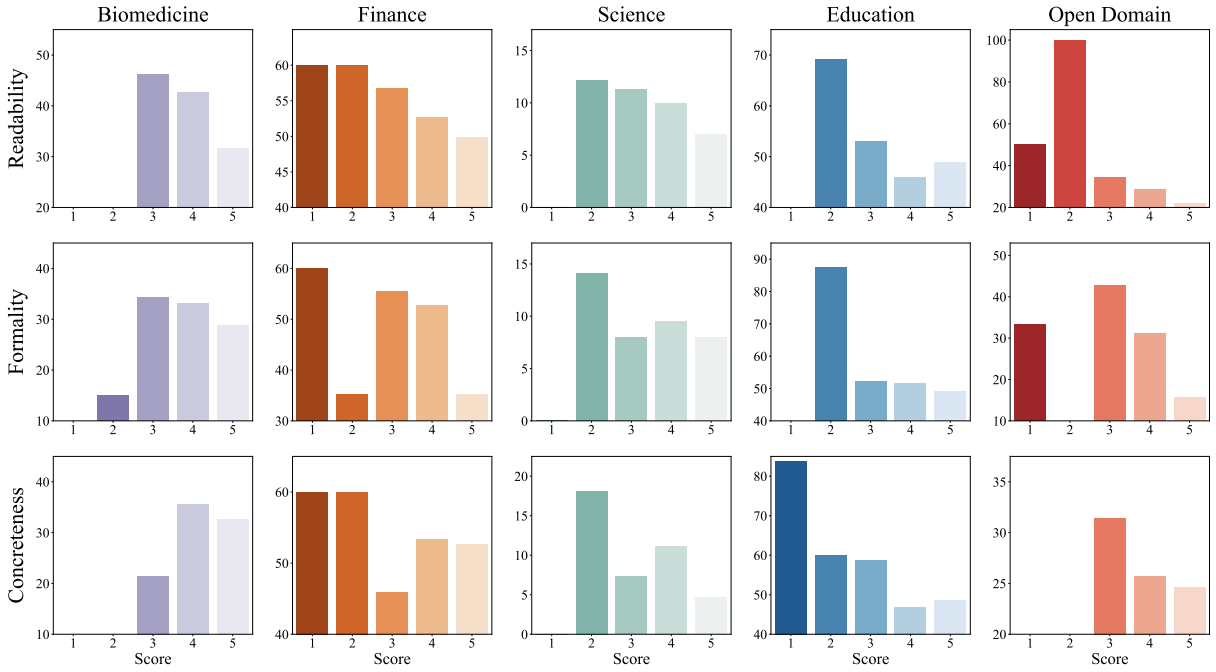
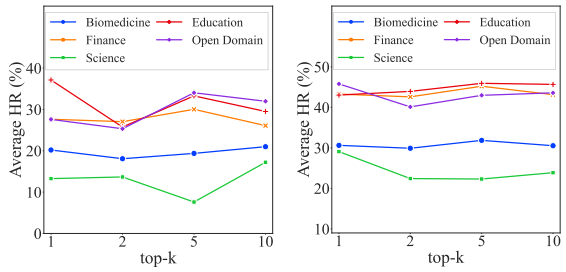
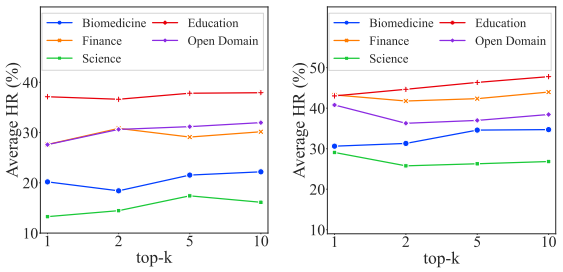


Figure 4: The average hallucination rate (%) of those responses and questions by ChatGPT for each score of the three properties, *i.e.,* readability, formality, and concreteness, in five domains. Some values are zero because there are no scores from humans. The 5-point denotes "very satisfying" and the 1-point denotes "very terrible".



(a) ChatGPT (b) Llama 2-Chat (7B)

Figure 5: Average hallucination rate (%) of using top-k retrieved documents as context.



(a) ChatGPT (b) Llama 2-Chat (7B)

Figure 6: Average hallucination rate (%) of sampling one document from top-k retrieved documents.

ments.

E.2 Self-Reflexion

Reflexion (Shinn et al., 2023) is an effective means through which LLMs can learn from and rectify their mistakes. Specially, the prior failure trials are transformed as textual feedback, which would be incorporated as additional context for the LLM itself. With the guidance of feedback, the LLM can generate an improved plan for the next attempt, called *self-reflexion*. Existing research (Ji et al., 2023b; Dhuliawala et al., 2023b) has demonstrated that self-reflexion is an effective method for hallucination mitigation. However, self-reflexion is a complex and advanced ability that requires mistake perception, feedback summarization, and behavioral planning. In light of this, we aim to explore to which extent self-reflexion can mitigate hallucinations and how the capacities of LLMs affect the reflexion performance on hallucination mitigation.

Experimental Details. We conduct self-reflexion experiments on Llama 2-Chat 7B, 13B, and 70B, to examine its effect on hallucination mitigation. Specifically, for a given query, we first obtain the initial model response, then require the model itself with instructions to reflect on whether its response contains any errors. If errors are detected, the model is prompted to provide a corrected response. This process is repeated until the model response is error-free or the maximum number of reflection iterations is reached. We set the maximum

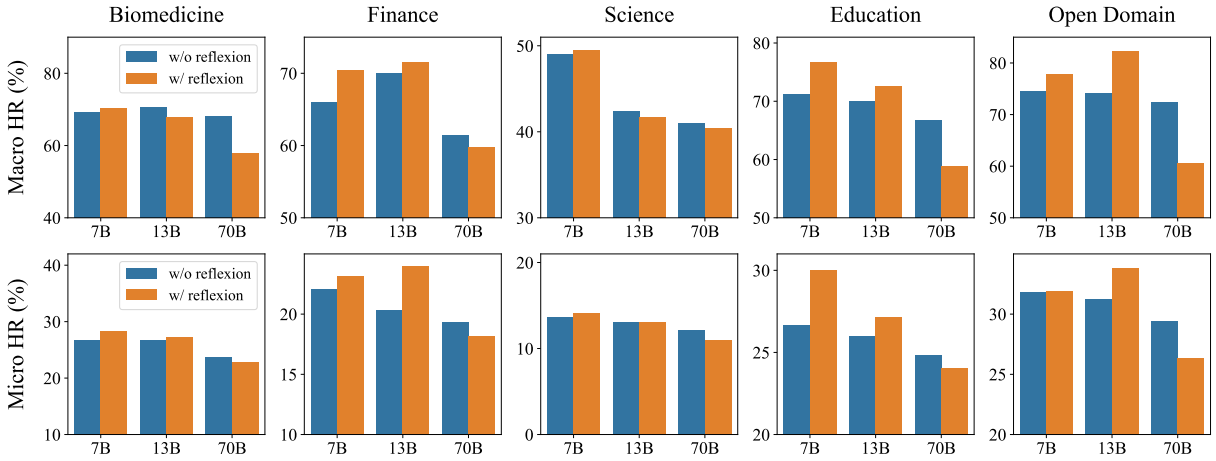


Figure 7: Macro and micro hallucination rate (%) for Llama 2-Chat (7B, 13B and 70B) with or without self-reflexion.

number of reflection iterations to 5. In addition, we also test these models without self-reflexion for performance comparison.

Results and Analysis. Figure 7 illustrates the relationship between the model's scale and its effectiveness in mitigating hallucinations through self-reflexion. We can clearly see that only when reaching a certain scale (*i.e.,* 70B in our experiments), the LLM possesses the self-reflective ability to mitigate hallucinations in its responses. For Llama 2-Chat 7B and 13B, self-reflexion, on the contrary, can even result in more erroneous responses. We found that due to the limited capacity of smaller models, the reflective behavior instead makes them suspect their original correct answers and generates wrong ones. Furthermore, the benefits obtained from self-reflexion are domain-sensitive, *e.g.,* self-reflexion reduces the LLM hallucinations slightly in the fields of finance and science, while significantly in the open domain.

E.3 Prompt Improvement

According to the results and analysis in Section D.3, we improve the task description, question expression, and in-context demonstrations in the original prompt. Furthermore, chain-of-thought (CoT) has been proven to be helpful in hallucination mitigation (Wang et al., 2023), so we explore adding CoT reasoning into our prompt.

Experimental Details. We improve the original base prompt from the following aspects:

- We incorporate *Base prompt*, *Manual description prompt*, and *Manual in-context prompt* from Section D.3 for comparison.
- *Domain info prompt*: injecting domain information into the task description in the base prompt.
- *Character role prompt*: defining a particular role (*e.g.,* scientist) for the system in the base prompt.
- *Zero-shot cot prompt*: adding zero-shot CoT to the base prompt by prepending "`Let's think step-by-step`".
- *Few-shot cot prompt*: adding few-shot CoT to the base prompt by injecting CoT examples.

Similarly, we test ChatGPT and Llama 2-Chat (7B) with these improved prompts to generate responses. For *few-shot cot prompt*, the few-shot reasoning examples are manually-written and different for each domain.

Results and Analysis. We show the prompt improvement results in Table 14. We can found that injecting domain information into task description or defining a character role for the system has an oscillatory effect on mitigating the hallucinations in LLM's responses. In the domains of finance and science, the two prompt improvement strategies can help LLMs generate more accurate responses. For Chain-of-Thought (CoT) prompting, its effect on hallucination mitigation heavily depends on the specific LLMs. For larger models like ChatGPT which possess exceptional reasoning capabilities, zero-shot or few-shot CoT reasoning can significantly benefit reducing the hallucinations. While for smaller models like Llama 2-Chat (7B) with limited reasoning abilities, leveraging CoT reasoning fails to effectively eliminate hallucinations and instead exacerbates the presence of hallucinations in their responses. Therefore, when engaging in prompt engineering, it is crucial to comprehensively consider factors such as model size, domain

Models	Biomedicine		Finance		Science		Education		Open Domain	
	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR	MaHR	MiHR
ChatGPT										
w/ base prompt	48.75	14.03	46.84	13.55	24.14	6.39	53.44	17.19	59.77	17.93
+ manual desc	45.64	13.91	39.20	11.18	22.34	5.28	55.68	17.73	64.52	20.31
+ manual demo	42.71	14.89	40.74	12.12	25.27	7.11	56.41	19.24	44.72	21.88
+ domain info	51.02	14.40	42.21	13.17	26.13	6.43	53.76	16.05	61.83	19.80
+ character role	52.50	14.67	44.67	12.72	26.13	6.33	54.30	17.11	60.85	19.37
+ zero-shot cot	46.60	13.51	41.33	12.35	24.62	6.26	53.11	15.96	56.90	17.21
+ few-shot cot	38.98	13.94	46.88	11.88	21.00	5.07	57.79	25.22	55.12	20.71
Llama 2-Chat 7B										
w/ base prompt	69.12	26.69	69.41	24.59	49.25	14.05	71.52	27.74	77.35	33.15
+ manual desc	68.02	26.46	74.36	25.01	42.50	12.10	76.16	30.97	79.39	33.23
+ manual demo	69.70	27.90	66.33	24.61	45.00	12.27	71.01	27.02	66.88	31.84
+ domain info	77.66	30.48	71.72	24.55	45.50	13.72	78.82	30.72	78.61	36.33
+ character role	73.23	32.12	73.87	25.60	47.50	13.42	73.21	27.67	84.62	35.79
+ zero-shot cot	77.84	30.34	78.61	27.03	50.25	15.99	79.61	30.77	69.12	31.20
+ few-shot cot	71.21	28.47	73.58	25.70	48.00	15.00	70.66	29.99	71.93	32.68

Table 14: Evaluation results of ChatGPT and Llama 2-Chat (7B) using different prompt improvement strategies.

characteristics, and task difficulty.