C-FAITH: A Chinese Fine-Grained Benchmark for Automated Hallucination Evaluation

Anonymous ACL submission

Abstract

Despite the rapid advancement of large language models, they remain highly susceptible to generating hallucinations, which significantly hinders their widespread application. Hallucination research requires dynamic and fine-grained evaluation. However, most existing hallucination benchmarks (especially in Chinese language) rely on human annotations, making automatical and cost-effective hallucination evaluation challenging. To address this, we introduce HaluAgent, an agentic framework that automatically constructs fine-grained QA dataset based on some knowledge documents. Our experiments demonstrate that the manually designed rules and prompt optimization can improve the quality of generated data. Using HaluAgent, we construct C-FAITH, a Chinese QA hallucination benchmark created from 1,399 knowledge documents obtained from web scraping, totaling 60,702 entries. We comprehensively evaluate 16 mainstream LLMs with our proposed C-FAITH, providing detailed experimental results and analysis.

1 Introduction

Despite significant advances made by large language models (LLMs) (Grattafiori et al., 2024; OpenAI et al., 2024) in natural language generation, hallucination continues to undermine their reliability and safety (Xu et al., 2024; Huang et al., 2024). The issue of hallucination makes the deployment of LLMs potentially risky in real-world applications (Bang et al., 2023; Ji et al., 2023). To understand what types of content and to what extent LLMs tend to hallucinate, much attention has been paid to constructing high-quality datasets for hallucination evaluation (Wang et al., 2023; He et al., 2024). Since the potential hallucinations of LLMs exist in various domains, the size and scalability of datasets are crucial for the oversight of LLM hallucinations.

However, constructing and scaling-up hallucination evaluation datasets face significant challenges

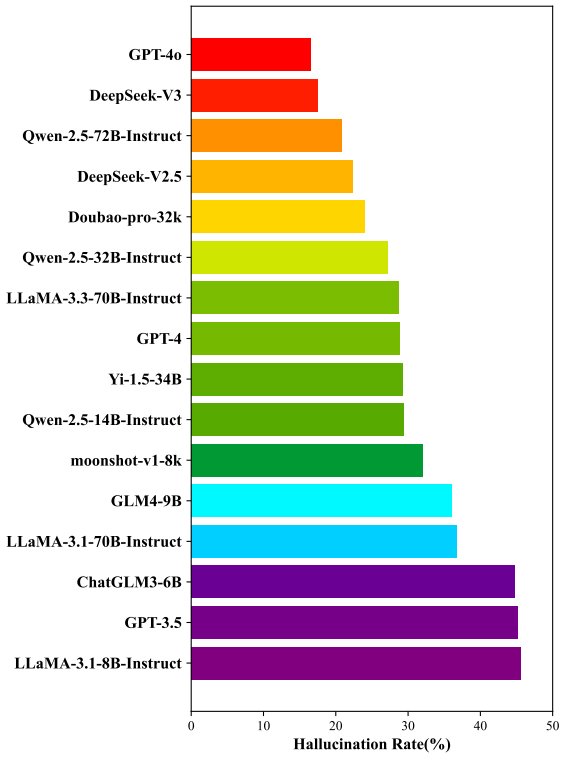


Figure 1: The total hallucination rates of 16 tested LLMs on C-FAITH.

(Cao et al., 2024; Liu et al., 2024b; Gu et al., 2024). As existing hallucination benchmarks (Li et al., 2023; Chen et al., 2024b) often rely on human annotations to construct high-quality datasets, one primary challenge is the prohibitively high costs of human annotation required for hallucination benchmark construction. Since manually constructing hallucination benchmarks is time-consuming and expensive, there is a need to develop automatic approaches to construct hallucination evaluation datasets at scale. To automate the construction of hallucination evaluation dataset, we propose **HaluAgent**, an agentic framework for automatic dataset generation.

We highlight key differences between existing hallucination benchmark construction and our

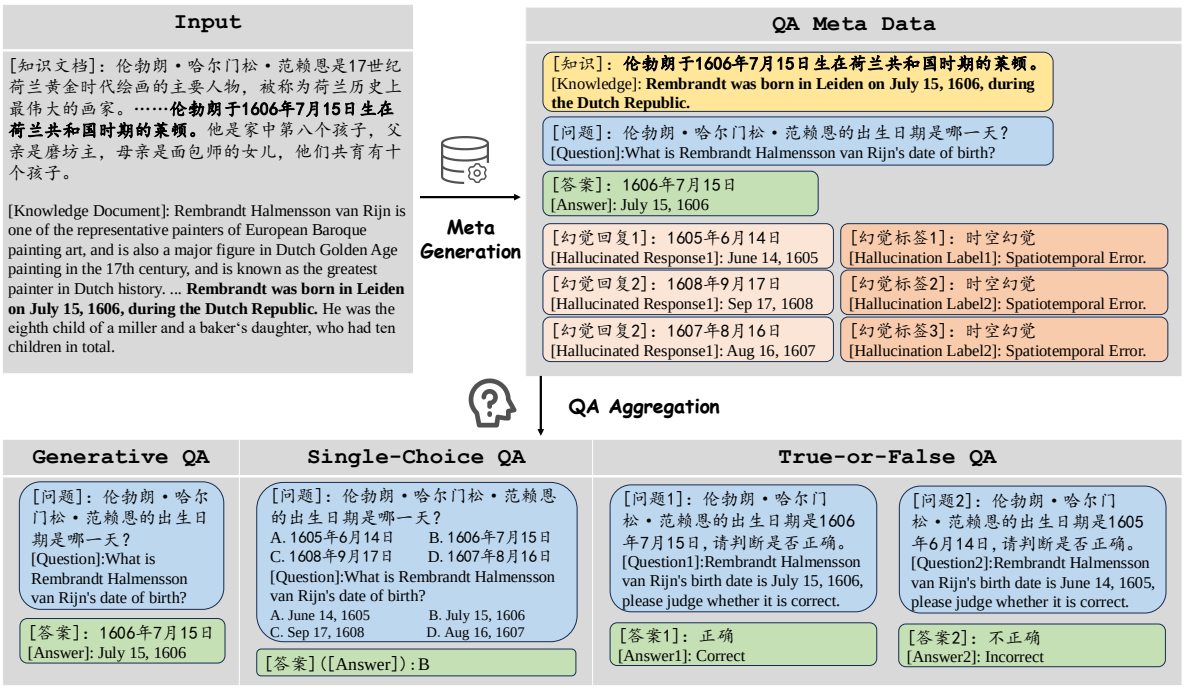


Figure 2: An example of the created QA data. HaluAgent first generates meta data containing question, correct answer, hallucinated responses and hallucination labels. Then, the QA meta data is aggregated into three different formats for hallucination evaluation. We provide both the Chinese QA data and the English translation in the figure.

HaluAgent method below:

- **Automatic data constuction:** Unlike some of the previous works that rely on manual annotation for data construction, we use a multi-agent system, HaluAgent. Built on Qwen model (Yang et al., 2024a), HaluAgent automatically generates and verifies QA data for hallucination evaluation.

- **Fine-grained error types:** Previous studies classify QA data based on question patterns (Chen et al., 2024b) or topic (Ji et al., 2024). Our study induces different types of hallucination given fine-grained hallucination error types. By generating hallucination labels for the QA data, HaluAgent enables more targeted identification of LLM hallucination.

- **Prompt optimization:** We leverage prompt optimization techniques (Yang et al., 2024b) to refine prompts for data generation. This improves the quality of data generation over methods that use simple few-shot prompting.

- **Multiple QA formats:** Previous studies often generate a single form of QA pairs, while HaluAgent supports multiple QA formats. Different forms of QA data help identify vulnerabilities of LLMs.

Given knowledge documents as input, HaluAgent supports the generation of three different QA data formats, including generative QA, single-choice QA and true-or-false QA. Figure 2 shows an example of generated data with a piece of knowledge extracted from the input document. Our HaluAgent method works by first generating questions and the corresponding correct answers. Then, we generate hallucinated responses for the questions, which are inconsistent with the background knowledge as well as the correct answers. The hallucinated response represents a potentially hallucinated generation. Finally, we produce hallucination labels for each hallucinated response to indicate the hallucination type induced by the question. By combining these elements, we obtain the QA meta-dataset. Three different QA formats can then be derived from this meta-dataset.

A verification module is proposed to check the correctness of the generated answers, the hallucinated responses, and the hallucination labels. During the large-scale data generation process, validated QA data are retained as the final hallucination evaluation dataset. Any data flagged as uncertain by the verification module is filtered out to ensure quality and accuracy. Moreover, to increase the validation rate of the generated data, we perform prompt optimization before large-scale generation.

As most existing hallucination benchmarks fo-

Factual Fabrication(**FactFab**)	The LLM fabricates concepts that do not exist or make up facts that do not exist in the real world.
Attribute Error(**AttrErr**)	The LLM generates incorrect content when describing object in the real world, such as the composition and function of an object.
Entity Error(**EntErr**)	The LLM generates text that contains false entities that contradict the world knowledge, such as people, event names, movies, books.
Relation Error(**RelErr**)	The LLM generates text containing false relationships between entities such as quantity, space, time, etc.
Spatiotemporal Error(**SpaErr**)	The LLM generates incorrect information about the time and space of an event.
Reference Error(**RefErr**)	The LLM makes up references and links that do not exist to make the generation more reliable.

Table 1: Classification of hallucinations in LLMs.

cus on English corpora, the number of QA datasets for Chinese hallucination evaluation is relatively limited. Therefore, we construct a hallucination benchmark to facilitate hallucination evaluation in the Chinese language in this paper. We collect 1,399 Chinese knowledge documents from multiple domains to construct a Chinese hallucination benchmark. Building on HaluAgent, we introduce **C-FAITH**, a benchmark comprising 16,713 generative QA items, 10,563 single-choice QA items and 33,426 true-or-false QA items in general. We evaluate 16 mainstream LLMs with C-FAITH, including both open-source and black-box models. Figure 1 summarizes the total halucination rates of 16 tested LLMs on our proposed C-FAITH. Moreover, we analyze the hallucination rates of LLMs when faced with different types of questions. The experimental results indicate that LLMs are most prone to hallucination involving entity errors and spatiotemporal errors.

In summary, our contributions can be listed as follows[1]:

- We propose **HaluAgent**, an automated framework for generating hallucination evaluation datasets with fine-grained error types in different formats.
- We introduce **C-FAITH**, a new Chinese hallucination evaluation dataset with fine-grained error types designed for the systematic assessment of hallucination generated by LLMs.
- We evaluate and analyze the hallucination risks of 16 mainstream LLMs with **C-FAITH**, providing detailed experimental results and analysis.

[1]Our code and data will be released to facilitate future research when accepted.

2 Related Work

Hallucination benchmarks construct challenging queries in single or multiple tasks to assess hallucination rate in LLM responses. These benchmarks cover a wide range of topics and tasks (Elaraby et al., 2023; Pal et al., 2023; Luo et al., 2024; Liu et al., 2024a). There are benchmarks curated with semi-automated approaches for data generation (Chen et al., 2024b,a; Ji et al., 2024; Mishra et al., 2024a) to offer better expandability compared with datasets that rely solely on manual annotations (Wang et al., 2023; Cheng et al., 2023). UHGEval (Liang et al., 2023) automates the construction of hallucination benchmarks for text continuation task. C-FAITH provides an automated, fine-grained and scalable hallucination benchmark for QA task.

Another line of work involves training a hallucination detector to evaluate the hallucination level of LLM generation (Muhlgay et al., 2023; Sriramanan et al., 2024; Gu et al., 2024; Akbar et al., 2024). Some early studies (Wang et al., 2020; Durmus et al., 2020; Liu et al., 2021; Dziri et al., 2022; Gupta et al., 2022; Laban et al., 2022; Varshney et al., 2023; Yang et al., 2023) focus on distinguishing whether the LLM output contains hallucinated content. Recent researches (Mishra et al., 2024b; Ji et al., 2024) further detect hallucinations in a more fine-grained and meticulous way.

3 Method

In this section we present our HaluAgent method. We use an agentic framework for data generation inspired by recent advances in role-playing (Park et al., 2023) and prompt optimization (Yang et al., 2024b). HaluAgent incorporates automated agents based on an open-source model Qwen-2-

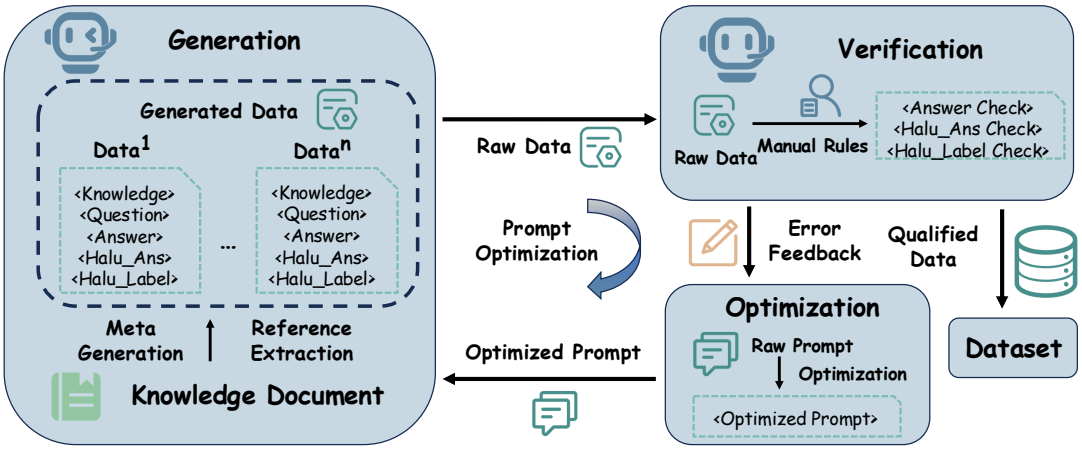


Figure 3: An illustration of our proposed HaluAgent framework. HaluAgent consists of three modules, including the generation module, the verification module and the optimization module. With manually designed rules, HaluAgent first conducts prompt optimization based on error feedbacks from the verification module. Next, HaluAgent takes knowledge documents as input to generate fine-grained QA data for hallucination evaluation.

72B-Instruct[2] (Yang et al., 2024a) to extract QA pairs from knowledge documents, generate hallucinated responses, and classify these hallucinated responses. Then, HaluAgent validates the correctness of the generated answers, hallucinated responses, and hallucination labels based on predefined manual rules. With the generated meta dataset, we aggregate it into different formats of QA data to construct the evaluation dataset.

To improve the quality of the generation data, we perform prompt optimization based on error feedbacks to refine the generation prompt before large-scale data generation. We then aggregate the generated data into multiple QA formats. Below, we discuss the key components of our HaluAgent.

3.1 Automatic Generation with Agentic Framework

Several recent works on hallucination benchmarks leverage LLMs to help create evaluation data. However, such methods (Li et al., 2023) require manual annotation to modify and validate LLM-generated outputs. As illustrated in Figure 3, our approach decomposes data construction into generation and verification modules. HaluAgent automatically generates QA data and verifies the correctness of the generated data with predefined strict manual rules.

By prompting the Qwen-2-72B-Instruct model, we build multiple data-generation agents and verification agents to generate and check the data. For each agent, the prompt structure begins with an initial instruction that specifies the task, followed by a set of rules outlining the task requirements. It also includes a selection of example inputs accompanied by their manually constructed outputs. Finally, the prompt ends with the target input for which the agent needs to generate a target output or check the given inputs. By providing this comprehensive prompt, we aim to teach the model to generate data according to the requirements and verify it based on the input rules.

The generation process is divided into three parts. Firstly, HaluAgent extracts knowledge from the knowledge document and generates the question and its correct answer. Then, HaluAgent generates hallucinated responses using the question and correct answer as input. Hallucinated responses are responses to the question that are inconsistent with the provided facts. Finally, HaluAgent takes the hallucinated responses and the correct answer as input to generate the hallucination label for each hallucinated response. We currently distinguish six types of LLM hallucination (Table 1). Different types of hallucinations are exemplified in Table 7 in the appendix. The generated hallucination label specifies the type of hallucination in the hallucinated responses.

To check the correctness of the generated QA data, we introduce the verification module within HaluAgent. Similar to the generation process, the verification process consists of three parts, namely correctness check, hallucination check and label check. Firstly, HaluAgent checks whether the correct answer is a valid response to the question based on the extracted knowledge. Then, HaluAgent checks whether the generated hallucinated

[2]Other LLMs can be used to construct the evaluation data as well.

responses are responses that contain hallucinated content. Finally, HaluAgent performs a label check to determine whether the hallucination label satisfies the definition of the hallucination type. We manually design verification rules for each hallucination type. If the generated data does not satisfy our proposed rules, we consider it as unqualified data. If HaluAgent is uncertain about the correctness of the input, it filters out such cases to ensure the quality of the generated data. Detailed rules are provided in Appendix A.

3.2 Prompt Optimization

Despite careful design, existing studies have shown that directly prompting LLMs (even GPT-4) to generate QA data still results in suboptimal quality (Long et al., 2024; Gu et al., 2024). We believe this might stem from the incomprehensive of the initial prompt and its stylistic inconsistency with the LLM. With meticulously designed requirements and a selection of examples for data generation, only 62.50% of the data generated by Qwen-72B model with the initial few-shot prompting is qualified in a sampled subset (Table 3). The low validation rate brings two major problems: 1) **Reduced data volume**: Due to the high rejection rate of generated data, the actual number of qualified evaluation samples we obtain is reduced. 2) **Increase of resource consumption**: A large amount of computing resources are wasted on the generation of unqualified data.

Therefore, improving the validation rate of generated data is critical for automating the process of data construction. We introduce prompt optimization to refine the prompt for data generation, making it more detailed and accurate, and better aligned with the LLM style. Our proposed prompt optimization follows a multi-round iterative framework. The optimization process samples a small number of knowledge documents to construct the training and validation datasets. We use the verification module to output error feedback for unqualified data and introduce an optimization agent to help optimize prompts. In each round of optimization, the optimization agent modifies the generation prompt according to the error feedback and generates a set of candidate optimized prompts. We select the optimal prompt for the next round of optimization based on the validation rate of the generated data conditioned on the candidate prompts.

	Corr.	**Halu.**	**Label**	α_{corr}	α_{halu}	α_{label}
W/o Ver	96.33%	92.00%	76.00%	0.87	0.79	0.82
W/Ver	98.67%	99.00%	94.33%	0.80	0.89	0.71

Table 2: Human annotation results on the verified QA data. **Corr.**, **Halu.** and **Label** denote the accuracy of correction check, hallucination check and label check given by human annotators respectively. We also provide α_{corr}, α_{halu} and α_{label}, which denote the Krippendorff's alpha (Krippendorff, 2004) of human annotation.

3.3 Aggregation for Multiple QA Formats

With the optimized prompt, we employ HaluAgent for large-scale data generation. For each input knowledge document, we first generate the raw data needed for constructing QA pairs, including the question, the correct answer, three possible hallucinated responses, and the hallucination labels. We construct questions for single-choice QA and true-or-false QA with the templates. In single-choice QA, we randomly shuffle the order of the options to prevent potential bias.

3.4 HaluAgent Results

To assess the effectiveness of HaluAgent, we conduct experiments to illustrate the contribution of the verification module and evaluate the impact of prompt optimization.

The verification module in HaluAgent effectively detects problematic data. To analyze the role of the verification module in HaluAgent, we compare the quality of the generated data with and without the verification module through human evaluation. Human annotators are recruited to label the correctness of the generated correct answers, the hallucinated responses, and the hallucination labels given the extracted knowledge. We randomly sample 300 generated correct answers, hallucinated responses, and hallucination labels from the verified dataset and the original dataset respectively for evaluation. We employ three human annotators for each sample. The human annotators are provided with the designed rules to guide their annotations. In cases of inconsistent annotations, we adopt a voting mechanism to select the majority label as the final human annotation result. Detailed instructions for human evaluation are provided in Appendix B. Table 2 provides the comparison of data correctness with and without the verification module. The verification module in HaluAgent effectively identifies errors in the generated data, thereby improving the accuracy evaluated by human annotation.

Iteration	Corr.↑	Halu.↑	Label↑	Overall↑
0	97.50%	88.00%	71.00%	62.50%
1	97.10%	95.65%	85.51%	81.16%
2	**98.55%**	**98.55%**	**88.41%**	**88.41%**
3	95.38%	92.31%	87.85%	86.15%

Table 3: The validation rate of generated data on the validation set after each round of prompt optimization. Corr., Halu. and Label denote the pass rate of correction check, hallucination check and label check respectively. The overall column illustrates the pass rate of all three checks.

Prompt optimization helps improve data generation quality. As human evaluation demonstrates the effectiveness of the verification module, we subsequently adopt the verification module to evaluate and enhance the performance of the generation module. We sample 60 and 20 knowledge documents from various domains as the training and validation inputs. We set the maximum number of prompt optimization iterations to five. If the validation rate on the validation set decreases during the optimization process, we terminate the prompt optimization early and select the optimal prompt on validation set as the final result. As illustrated in Table 3, we provide the performance of the generated data at different iteration stages during the prompt optimization process. Through two rounds of iteration, the generation prompt achieves the best performance on the validation set, with an overall success rate of 88.41%, surpassing the initial result by 25.91%.

Optimized prompt enhances the clarity and details of the prompt. To investigate how HaluAgent modifies the prompt during the prompt optimization process, we manually examine prompt cases during optimization. Figure 4 presents the original and optimized prompts for defining spatiotemporal error when generating hallucination labels. Prompt optimization adds more details and requirements to the generation process, making the objectives more clear. In addition, the optimization incorporates examples into the generated content to enhance the clarity of the explanations. The complete results of prompt optimization are presented in Figure 7 in the Appendix.

4 C-FAITH: A Chinese Fine-Grained Question Answering Benchmark

Using our HaluAgent method, detailed in Section 3, we introduce C-FAITH, a Chinese fine-grained QA hallucination benchmark. By converting knowledge documents from various domains into question answering evaluation data, C-FAITH covers multiple domains. This benchmark provides a scalable resource for Chinese hallucination evaluation with fine-grained error types.

Original prompt

时空幻觉:问题是对时间或空间提问，正确答案和幻觉回复都是时间或空间概念。
Spatiotemporal Error: The question asks about time or space, with the correct answer and the hallucinated response containing correct and incorrect time or space concepts respectively.

Optimized prompt

时空幻觉:问题是对时间或空间提问，正确答案和幻觉回复都是时间或空间概念，但回复中的时间或空间信息与正确答案不符。例如，当询问某个事件发生的时间时，幻觉回复给出了一个明显错误的时间点。
Spatiotemporal Error: The question asks about time or space, with the correct answer and the hallucinated response containing correct and incorrect time or space concepts respectively. For example, when asking about the time an event occurred, the hallucinated response provides a clearly incorrect timestamp.

Figure 4: The prompt optimization case of spatiotemporal error definition in hallucination label generation.

Overall	Gen.	Choice.	True/False.
60,702	16,713	10,563	33,426

Table 4: Data Statistics of our proposed C-FAITH.

4.1 Dataset Overview

We retrieve 1,399 knowledge documents from several Chinese databases including Wikipedia-CN[3], Baidu Baike[4], Sougou Baike[5]. These documents cover six major topics: Celebrities, Entertainment, Education, Astogeography, Biology, and Culture. We categorize the hallucination type of a question based on the primary hallucination category of the generated hallucinated responses. This hallucination type represents the specific type of hallucination the question is designed to induce[6]. HaluAgent generates various types of questions from each knowledge document to elicit different hallucinations. Since hallucinated responses to the same question may be highly similar, we perform deduplication for single-choice QA. Only single-choice QA data containing distinct hallucinated options

[3]https://zh.wikipedia.org/
[4]https://baike.baidu.com/
[5]https://baike.sogou.com/
[6]This hallucination type is the type that LLMs are most likely to generate and can also be referred to as the primary hallucination type.

Model	Gen. ↓	Choice. ↓	True/False. ↓	Ova. ↓
GPT-4o	25.17	**10.40**	**14.30**	**16.53**
GPT-4	43.43	22.58	22.85	28.88
GPT-3.5	45.69	42.60	38.70	45.19
Doubao-pro-32k	34.69	17.42	22.51	23.96
moonshot-v1-8k	44.42	25.48	26.93	31.99
DeepSeek-V2.5	29.65	17.62	19.35	22.31
DeepSeek-V3	**24.25**	11.60	16.70	17.52
ChatGLM-6B	65.79	37.30	43.90	44.80
GLM4-9B	50.55	31.10	33.80	36.01
LLaMA-3.1-8B	67.75	41.60	33.15	45.62
LLaMA-3.1-70B	56.69	19.30	28.25	36.79
LLaMA-3.3-70B	44.88	19.20	23.20	28.77
Yi-1.5-34B	43.26	23.70	24.90	29.26
Qwen-2.5-14B	58.06	19.60	24.85	29.46
Qwen-2.5-32B	41.22	17.90	21.25	27.13
Qwen-2.5-72B	32.63	14.40	16.40	20.82

Table 5: The hallucination rate(%) of the generated content from various LLMs. We calculate the hallucination rate of LLMs across the three formats of QA data and computed their average as the overall hallucination rate. We bold the best-performing model and underline the second-best-performing model.

are retained. The statistics and distribution of the C-FAITH dataset are shown in Table 4 and Figure 6, including the distribution of topics and hallucination types for the questions in C-FAITH. Finally, we evaluate various LLMs using the proposed C-FAITH dataset.

4.2 Experimental Settings

Models We evaluate 16 LLMs with C-FAITH, including GPT-4o, GPT-4 (OpenAI et al., 2024), GPT-3.5(Brown et al., 2020), Doubao-pro-32k, moonshot-v1-8k, DeepSeek-V2.5, DeepSeek-V3 (DeepSeek-AI, 2024), ChatGLM3-6B (Du et al., 2022), GLM4-9B (GLM et al., 2024), LLaMA-3.1-8B-Instruct, LLaMA-3.1-70B-Instruct, LLaMA-3.3-70B-Instruct (Grattafiori et al., 2024), Yi-1.5-34B (AI et al., 2024), Qwen-2.5-14B-Instruct, Qwen-2.5-32B-Instruct and Qwen-2.5-72B-Instruct (Qwen-Team, 2024). We adopt the same decoding settings (temperature=1, top_p=0.7) for all LLMs.

Metrics The evaluation is conducted in the zero-shot setting. We calculate the hallucination rate of LLM generations as the evaluation metric. For generative QA, to evaluate whether the LLM output contains hallucination, we employ GPT-4o to determine if the LLM output conflicts with the correct response. If a conflict is identified, the LLM output is considered to be hallucinated. For single-choice QA and true-or-false QA, we determine the correctness of the LLM output with the correct answer by directly comparing it with the correct answer. We calculate the average hallucination rate across the three QA formats as the overall hallucination rate for each LLM.

4.3 Results and Analysis

Table 5 presents the experimental results for 16 different LLMs. In general, GPT-4o shows the lowest overall hallucination rate, followed by DeepSeek-V3 and Qwen-2.5-72B-Instruct. Deepseek-v3 performs well in the the generative QA task achieving a lower hallucination rate than GPT-4o. Among the various QA formats, LLMs are more prone to generating hallucinations in generative QA tasks, leading to lower response accuracy. In most cases, the models exhibit similar relative performance across the three QA formats. However, some LLMs, such as LLaMA-3.1-70B, show low accuracy in generative QA but perform well in single-choice QA, even achieving higher accuracy than GPT-4.

In addition, we observe the following insights:

Larger LLMs generally exhibit lower overall hallucination rate. As expected, scaling up model sizes typically leads to low hallucination rates. Under the same model architecture, such as in the Qwen-2.5 series and LLaMA3 models, the hallucination rate decreases as the model's parameter size increases across all three QA formats.

A significant difference exists in hallucination rate across different hallucination types. As C-FAITH provides the hallucination label for each question, we analyze the impact of hallucination label on the hallucination rate of LLM response. We focus on experimental results of generative QA data. Table 6 summarizes the hallucination rate of LLM responses conditioned on different hallucination labels. LLMs have their own strengths when responding to questions of different hallucination labels. For example, the Qwen series achieve low hallucination rate when addressing quesitions related to attribute errors, while performimg poorly when handling spatiotemporal-related questions.

For the vast majority of LLMs, the probability of hallucinated generation is high when faced with questions meant to induce entity errors and spatiotemporal errors. On one hand, this is because the answers to these types of questions are deterministic and unique. Therefore, any discrepancy between the LLM generation and the facts results in hallucination. On the other hand, LLMs are generally prone to entity confusion and erroneous memory of time and location information.

Model	Overall↓	FactFab↓	AttrErr↓	EntErr↓	RelErr↓	SpaErr↓	RefErr↓
GPT-4o	25.17	**22.43**	13.01	31.77	**15.15**	**35.68**	**10.00**
GPT-4	43.43	38.82	24.49	55.71	30.30	54.92	30.00
GPT-3.5	54.31	52.96	33.61	67.95	30.30	65.36	30.00
Doubao-pro-32k	34.69	44.44	21.43	38.61	35.71	40.89	30.00
moonshot-v1-8k	44.42	45.20	25.04	53.71	36.36	57.55	30.00
DeepSeek-V2.5	29.65	30.96	14.62	35.30	18.18	41.97	23.33
DeepSeek-V3	**24.25**	24.69	**10.98**	**27.49**	18.18	38.96	16.67
ChatGLM3-6B	65.79	67.49	50.50	72.80	42.42	77.98	50.00
GLM4-9B	50.55	53.25	29.03	59.89	36.36	65.80	36.67
LLaMA-3.1-8B-Instruct	67.75	65.33	50.17	78.43	42.42	80.05	50.00
LLaMA-3.1-70B-Instruct	56.69	55.73	41.95	65.38	60.61	64.77	36.67
LLaMA-3.3-70B-Instruct	44.88	41.80	27.35	56.04	30.30	55.70	26.67
Yi-1.5-34B	43.26	41.18	22.32	55.08	33.33	56.22	33.33
Qwen-2.5-14B-Instruct	41.94	41.80	19.63	51.99	39.39	59.07	20.00
Qwen-2.5-32B-Instruct	41.22	39.44	20.34	51.92	36.36	55.96	24.14
Qwen-2.5-72B-Instruct	32.63	33.75	16.11	39.42	27.27	45.85	13.33

Table 6: The hallucination rates(%) of LLM responses across different hallucination types on generation QA.

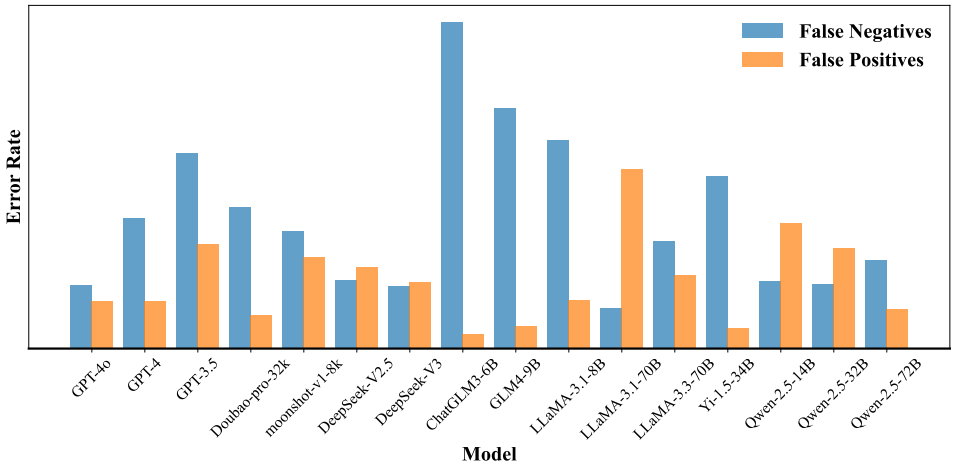


Figure 5: The false negative and false positive error rates of LLMs when facing true-or-false questions.

Through fine-grained hallucination classification, we help identify the main category of hallucination in LLMs, providing guidance for targeted hallucination mitigation strategies.

Most LLMs are more likely to be fooled by the input hallucinated content. For true-or-false QA, we categorize LLM errors into two types: false negative (where the model incorrectly classifies a hallucinated response as correct) and false positive (where the model incorrectly classifies a correct response as hallucinated). In Figure 5, we present a comparison of false negative and false positive error rates across different LLMs. Except for a few LLMs such as LLaMA-3.1-70B, most models exhibit a higher probability of generating false negatives than false positives. Models such as ChatGLM-6B and LLaMA3.1-8B exhibit a significantly higher false negative error rate compared to the false positive error rate. Even GPT-4 exhibits a clear bias to generate false negative errors. This phenomenon suggests that most LLMs tend to accept the input content as valid, even when the input itself contains hallucinations.

5 Conclusion

We address the limitations in automated fine-grained question answering benchmarks for hallucination evaluation by introducing HaluAgent, a multi-agent system that automatically constructs hallucination benchmarks. HaluAgent effectively creates questions, correct answers, hallucinated responses and hallucination labels according to the input knowledge document. Building on this, C-FAITH provides a benchmark of 60,702 QA data in total, providing a new fine-grained scalable benchmark for LLM hallucination evaluation.

Limitations

Currently, the C-FAITH evaluation dataset encompasses six types of hallucinations. We plan to refine and correct any misclassifications to ensure a more comprehensive evaluation of hallucinations. In addition, our dataset primarily focuses on six general knowledge domains. We aim to expand the evaluation to include other fields, such as healthcare and finance.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. 2024. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037, Miami, Florida, USA. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zouying Cao, Yifei Yang, and Hai Zhao. 2024. Autohall: Automated hallucination dataset generation for large language models. *Preprint*, arXiv:2310.00259.

Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. 2024a. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2403.00896*.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Jiang Yong, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2024b. Factchd: Benchmarking fact-conflicting hallucination detection. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. Anah-v2: Scaling analytical hallucination annotation of large language models. *Preprint*, arXiv:2407.04693.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu

Guo, Chengwei Hu, Boren Zheng, Zhuoran Lin, Xuepeng Liu, Dekai Sun, Shirong Lin, Zhicheng Zheng, Xiaoyong Zhu, Wenbo Su, and Bo Zheng. 2024. Chinese simpleqa: A chinese factuality evaluation for large language models. *Preprint*, arXiv:2411.07140.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. ANAH: Analytical annotation of hallucinations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

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

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Xun Liang, Shichao Song, Simin Niu, Zhiyu Li, Feiyu Xiong, Bo Tang, Yezhaohui Wang, Dawei He, Peng Cheng, Zhonghao Wang, et al. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296*.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. 2024a. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *Preprint*, arXiv:2403.11116.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Wen Luo, Tianshu Shen, Wei Li, Guangyue Peng, Richeng Xuan, Houfeng Wang, and Xi Yang. 2024. Halludial: A large-scale benchmark for automatic dialogue-level hallucination evaluation. *arXiv preprint arXiv:2406.07070*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024a. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024b. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

Qwen-Team. 2024. Qwen2.5: A party of foundation models.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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. *Preprint*, arXiv:2307.03987.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Binjie Wang, Ethan Chern, and Pengfei Liu. 2023. Chinesefacteval: A factuality benchmark for chinese llms.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *Preprint*, arXiv:2401.11817.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2024b. Large language models as optimizers. *Preprint*, arXiv:2309.03409.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3898–3908, Singapore. Association for Computational Linguistics.

A Manual rules in the verification module

We provide the manual rules used in the verification module in this section.

B Human Evaluation

We provide our human evaluation guideline furnished to participants for manually evaluating the correctness of the generated correct answer, hallucinated response and hallucination label. We recruited three Chinese college students for annotation.

Human Evaluation Instructions

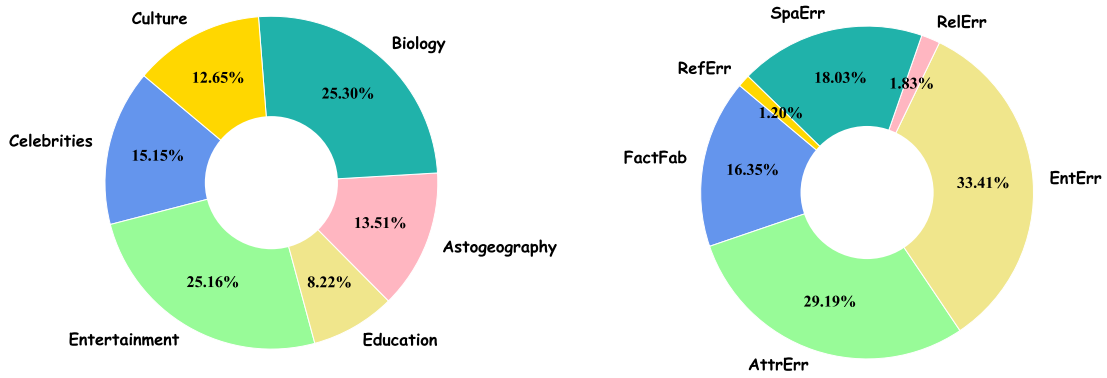
感谢参与本次标注任务！我们目前正在进行一个幻觉研究项目，需要评价生成幻觉评测数据的正确性。我们生成的数据包含问题、正确答案、背景知识、幻觉答复和幻觉标签。我们需要您评价正确答案正确性（要求：正确答案是对问题的正确回答，背景知识可以视为正确的事实）、幻觉回复的幻觉性（要求：幻觉答案中包含与背景知识不一致的幻觉的表述）、幻觉标签正确性（要求：幻觉标签与给出的幻觉标签的定义相符合）。具体的幻觉标签的定义如下：
虚构事实：[幻觉回复]必须包含不存在的概念或无法被现有资料证实的伪事实。[幻觉回复]中包含的内容在现实世界中没有实际依据，或者是错误地捏造出来的。例如，完全虚构的事件、人物、地点或其他无法查证的事实。如果[问题]中明确要求描述事物的某些属性，如功能、特征、组成等时，一律归为属性错误而不符合虚构事实的要求。
属性错误：[问题]中明确要求描述事物的某些属性，如功能、特征、组成等，而[幻觉回复]中包含对这些实体属性的错误表述。
实体错误：[问题]应针对具体实体进行提问。实体指的是特定的、具有实际存在或明确身份的事物，例如人名、地名、作品名、事件名、概念等。[幻觉回复]包含与[正确答案]或已知世界知识相矛盾的错误实体，例如人名、事件名、书籍、电影等。实体错误专注于对已知实体的错误描述。实体通常是现实世界中已经明确存在的东西，所以这个错误不涉及虚构内容，而是对已经存在的事物进行错误的归类或描述。
时空幻觉：[幻觉回复]必须涉及时间、空间或特定时期的描述。时空幻觉仅限于时间或地点描述错误，不包括对事件本身的内容或属性的误解。如果对事件本身内容表述有误，归类于虚构事实或属性幻觉。
关系错误：[幻觉回复]必须包含对不同实体之间的比较，而不是单独对每个实体的描述。[幻觉答案]与[正确答案]在描述不同实体之间关系时，必须存在不一致。
虚假引用：[幻觉回复]中必须包含虚假的引用或链接。

Manual rules for verification

正确答案检测：
假设你是一个中文问答正确性检验者。请根据给定的背景知识，判断所给答案是否是问题的正确回复并给出理由。具体对问答的要求如下：
1. [问题]必须能够在[背景知识]中找到正确答案。
2. [正确答案]必须是对问题的正确回复。
3. [正确答案]必须与背景知识相符，背景知识可以视为正确的事实。
4. [正确答案]自身必须符合事实和逻辑，足够合理。
<例子>
幻觉回复检测：
假设你是一个中文问答错误性检验者。请根据给定的背景知识，判断所给幻觉答案是否是问题的错误回复并给出理由。具体对问答的要求如下：
1. [幻觉答案]必须是对问题的错误回复。
2. [幻觉答案]必须与背景知识相违背，或与背景知识中的逻辑和信息不符。
3. [幻觉答案]自身必须符合事实和逻辑，足够合理。
<例子>
幻觉标签检测： 虚构事实：假设你是一个幻觉类型检测员。请根据给定的要求判断是否存在虚构事实，并提供理由。具体要求如下：
1. [幻觉答案]必须包含不存在的概念或无法被现有资料证实的伪事实。[幻觉答案]中包含的内容要么是完全不存在的概念，要么是无法通过现有可靠资料证实的伪事实。这些内容在现实世界中没有实际依据，或者是错误地捏造出来的。例如，完全虚构的事件、人物、地点或其他无法查证的事实。
2. [幻觉答案]与[正确答案]在事实或概念的表述上存在不一致，必须包含完全虚构的内容，与现实完全不符。[幻觉答案]中的内容应与[正确答案]在事实或概念上的表述完全不一致。这个不一致不是简单的误解或错误，而是包含了完全虚构的部分，和现实世界或公认的事实有根本的区别。例如，错误地描述一个不存在的历史事件，或提供一个没有任何依据的虚构数据。
3. [问题]中明确要求描述事物的某些属性，如功能、特征、组成等时，一律归为属性错误而不符合虚构事实的要求。
属性错误：假设你是一个幻觉类型检测员。请根据给定的要求判断是否存在属性错误，并提供理由。具体要求如下：
1. [问题]明确要求描述事物的某些属性，如功能、特征、组成等。[问题]本身会清楚地要求对某个具体实体或事物的属性进行描述。例如，问题可能询问某个物体的功能、组成、外观特征、用途、资格等。[问题]并不要求实体的整体或身份描述，而是聚焦于对该事物的某些具体属性的说明。
2. [幻觉答案]中包含对现实世界存在的物体进行错误的属性描述，通常是误导性的或完全不符合该事物的实际特性。[幻觉答案]将现实世界中的某个物体或事物的属性描述错误。这些错误的属性描述通常表现为错误的功能描述、错误的外观特征描述、错误的组成描述等。
3. [幻觉答案]与[正确答案]对实体属性的描述存在不一致。这种差异不仅仅是表述上的不一致，而是指对该事物核心属性的描述发生了根本错误或误解。例如，描述某个物体的特征时，幻觉答案的描述与正确答案显著不符，导致用户得到一个错误的认知。
实体错误：假设你是一个幻觉类型检测员。请根据给定的要求判断是否存在实体错误，并提供理由。具体要求如下：
1. [问题]应针对具体实体进行提问。实体指的是特定的、具有实际存在或明确身份的事物，例如人名、地名、作品名、事件名、概念等。[问题]应聚焦于这些实际存在的具体对象，而不是抽象概念。
2. [幻觉答案]包含与[正确答案]或已知世界知识相矛盾的错误实体，例如人名、事件名、书籍、电影等。实体错误专注于对已知实体的错误描述。实体通常是现实世界中已经明确存在的东西，所以这个错误不涉及虚构内容，而是对已经存在的事物进行错误的归类或描述。例如，如果问题问某部电影的导演，而幻觉答案给出了错误的导演名字，尽管电影确实存在，错误的名字就构成了实体错误。
时空幻觉：假设你是一个幻觉类型检测员。请根据给定的要求判断是否存在时空幻觉，并提供理由。具体要求如下：
1. [幻觉答案]必须涉及时间、空间或特定时期的描述。[幻觉答案]需要明确描述与时间（如年份、日期、时刻等）或空间（如地点、地理位置等）相关的内容。这种描述通常与事件发生的时刻、地点、持续时间等信息密切相关。比如提到某事件发生的具体年份、发生地点或涉及的时间范围。
2. [幻觉答案]与[正确答案]在时间或空间的描述上存在不一致。[幻觉答案]中的时间或空间信息必须与正确答案的实际时间或空间信息不符。这种不一致通常表现为错误的年份、日期、时刻，或是错误的地点描述。
3. 时空幻觉仅限于时间或地点描述错误，不包括对事件本身的内容或属性的误解。[幻觉答案]中对事件或属性描述的具体内容必须正确，只有时间或空间部分出错。
关系错误：假设你是一个幻觉类型检测员。请根据给定的要求判断是否存在关系错误，并提供理由。具体要求如下：
1. [幻觉答案]必须包含对不同实体之间的比较，而不是单独对每个实体的描述。
2. [幻觉答案]与[正确答案]在描述不同实体之间关系时，必须存在不一致。
虚假引用：假设你是一个幻觉类型检测员。请根据给定的要求判断是否存在虚假引用，并提供理由。具体要求如下：
1. [幻觉答案]中必须包含引用或链接。
2. [幻觉答案]中的引用或链接必须是虚构的，实际上这些引用和链接并不存在，或与[正确答案]中给出的引用和链接不一致。

Hallucination type	Question	Hallucinated Response
FactFab	Tell me about the historical origins of unicorns.	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.
AttrErr	Please introduce the function of a pen.	Pens are primarily used for painting. They are usually made of metal, and do not use ink.
EntErr	Which film won the Palme d'Or at the 75th Cannes Film Festival?	The Palme d'Or at the 75th Cannes Film Festival was awarded to the French film The Noon Star.
RelErr	Who was born first, Aaron Gillespie or Nathan Leone?	Aaron Gillespie was born before Nathan Leone.
SpaErr	What is Rembrandt Halmensson van Rijn's date of birth?	June 14, 1605.
RefErr	Give me five of the most influential research articles on large model hallucinations.	Towards a Rigorous Science of Neural Language Models" by LeCun et al. (which actually does not exist)

Table 7: Examples of specific subtypes of hallucinations.



(a) Topic distribution of C-FAITH

(b) Hallucination type distribution of C-FAITH

Figure 6: General Overview of C-FAITH dataset, containing topic and hallucination type distribution of the dataset.

Original rules for question generation

1. 生成的问题必须尽可能难。
2. 生成的问题必须与背景知识相关。
3. 生成的答案必须是生成问题的正确回复。
4. 生成的答案符合事实和逻辑。
1. The generated questions must be as difficult as possible.
2. The generated questions must be related to the background knowledge.
3. The generated answers must be the correct responses to the generated questions.
4. The generated answers must be factually accurate and logically sound.

Optimized rules for question generation

1. 生成的问题必须尽可能难，但应确保答案能在给定的背景知识中找到依据。
2. 生成的问题必须与背景知识相关，避免引入无关信息导致答案的不确定性。
3. 生成的答案必须是生成问题的正确回复，且应基于背景知识中的具体信息，避免假设和常识性的推断。
4. 生成的答案必须符合事实和逻辑，不能与常识相悖。
1. The generated questions must be as challenging as possible while ensuring that their answers can be substantiated by the provided background knowledge.
2. The generated questions must be relevant to the background knowledge, avoiding the introduction of extraneous information that could lead to answer uncertainty.
3. The generated answers must be the correct responses to the corresponding generated questions and should be based on specific information from the background knowledge, avoiding assumptions and commonsense-based inferences.
4. The generated answers must be factually and logically consistent, without contradicting common sense.

Original rules for hallucinated response

1. 生成的幻觉回复必须与问题相关，是一个问题的回复。
2. 生成的幻觉回复必须与依据不一致。
3. 生成的幻觉回复必须与正确答案不一致。
4. 生成的幻觉回复只需要包含回复本身，不能包含解释。
1. The generated hallucinated response must be relevant to the question and serve as a response to it.
2. The generated questions must be related to the background knowledge.
3. The generated answers must be the correct responses to the generated questions.
4. The generated answers must be factually accurate and logically sound.

Optimized rules for hallucinated response

1. 生成的幻觉回复必须与问题相关，是一个问题的回复，同时应巧妙地结合相关但非直接关联的信息以增加迷惑性。
2. 生成的幻觉回复必须与依据不一致，但应巧妙地融入部分真实信息以增加迷惑性。
3. 生成的幻觉回复必须与正确答案不一致，特别是对于单项得分要求的部分，可以提出一些看似合理但实际上与常规考试标准相违背的条件，同时确保这些条件在背景知识中没有直接反驳。
4. 生成的幻觉回复只需要包含回复本身，不能包含解释。同时，应确保其表述方式看似合理，能够吸引不了解背景知识的人信以为真，并在可能的情况下引用一些虚构或误解的数据、研究或权威观点来加强说服力。
1. The generated hallucinated response must be relevant to the question and serve as a response, while subtly incorporating related but not directly relevant information to enhance misleading potential.
2. The generated hallucinated response must be inconsistent with the given background knowledge but should skillfully integrate partial truths to increase plausibility.
3. The generated hallucinated response must differ from the correct answer. Specifically, for components requiring individual scoring, it may introduce seemingly reasonable conditions that actually contradict standard assessment criteria, while ensuring that these conditions are not explicitly refuted in the background knowledge.
4. The generated hallucinated response should contain only the response itself, without any explanation. Additionally, its wording should appear reasonable enough to convince those unfamiliar with the background knowledge. Where possible, it may reference fabricated or misinterpreted data, research, or authoritative opinions to enhance its persuasiveness.

Original prompt for hallucination label

虚构事实：幻觉回复犯了事实编造的错误。
属性错误：问题询问某事物的特征，幻觉回复对相关信息的介绍有误。
实体错误：正确答案是一个实体，并且幻觉回复是错误的实体。
时空幻觉：问题是对时间或空间提问，正确答案和幻觉回复都是时间或空间概念。
关系错误：问题必须要求对不同实体进行比较，而不是对实体分别提问，幻觉回复中也必须包含错误的实体关系。
虚假引用：幻觉回复引用名人名言、书籍或是网站等来源的知识佐证回复，但这引用的知识是错误的，与事实不一致。
Factual Fabrication: The hallucinated response contains factual fabrications.
Attribute Error: The question asks about the attribute of something and the hallucinated response provides incorrect information in the introduction.
Entity Error: The correct answer is an entity, but the hallucinated response presents an incorrect entity.
Spatiotemporal Error: The question pertains to time or space, and both the correct answer and the hallucinated response involve temporal or spatial concepts.
Relation Error: The question must require a comparison between different entities rather than asking about them separately, and the hallucinated response must contain incorrect entity relationships.
Reference Error: The hallucinated response references knowledge from sources such as famous quotes, books, or websites to support its claim, but the cited knowledge is incorrect and inconsistent with facts.

Optimized prompt for hallucination label

虚构事实：幻觉回复犯了事实编造的错误，即提供了与现实世界不相符的信息。
属性错误：问题询问某事物的特征，幻觉回复对相关信息的介绍有误，尤其是对现实世界存在的物体或概念进行错误的属性描述。如果讨论的是抽象概念（如考试要求）而非具体物体的属性，那么错误描述应与该抽象概念的本质特征相悖。当回答中夸大了某些特征的独特性和难度，即使这些特征确实存在，但与背景知识或常识中的普遍情况不符时，也属于属性错误。
实体错误：正确答案是一个实体，并且幻觉回复是错误的实体，即问题要求回答某一特定实体，而回复中提供了另一个不相关的实体。例如，当被问及某个具体人物时，回复提到了一个完全不同的个人。
时空幻觉：问题是对时间或空间提问，正确答案和幻觉回复都是时间或空间概念，但回复中的时间或空间信息与正确答案不符。例如，当询问某个事件发生的时间时，幻觉回复给出了一个明显错误的时间点。
关系错误：问题必须要求对不同实体进行比较，而不是对实体分别提问，幻觉回复中也必须包含错误的实体关系。在进行比较时，不仅需要识别出比较对象之间的关系描述是否准确，还需判断这种比较是否基于正确的信息基础之上。
虚假引用：幻觉回复引用名人名言、书籍或是网站等来源的知识佐证回复，但这引用的知识是错误的，与事实不一致。需要确认引用来源的真实性和准确性。
Factual Fabrication: The hallucinated response contains fabricated information that does not align with reality.
Attribute Error: The question asks about the attribute of something, but the hallucinated response provides incorrect information. This includes misattributing properties to real-world objects or concepts. If the discussion pertains to abstract concepts (e.g., examination requirements) rather than physical attributes, the error should contradict the fundamental nature of that concept. Additionally, exaggerating the uniqueness or difficulty of certain features—despite their existence—constitutes an attribute error if such claims contradict background knowledge or common understanding.
Entity Error: The correct answer is a specific entity, but the hallucinated response presents an incorrect entity. This occurs when the question requires a response about a particular entity, but the reply introduces an entirely unrelated one. For example, if asked about a specific historical figure, the response mistakenly mentions a completely different person.
Spatiotemporal Error: The question pertains to time or space, and while both the correct answer and the hallucinated response involve temporal or spatial concepts, the response provides incorrect temporal or spatial information. For instance, when asked about the date of an event, the hallucinated response gives an obviously incorrect timeframe.
Relation Error: The question must involve a comparison between different entities rather than addressing them separately. The hallucinated response must contain incorrect relationships between entities. When making comparisons, it is necessary to verify whether the description of relationships is accurate and whether the comparison is based on a correct factual foundation.
Reference Error: The hallucinated response references knowledge from sources such as famous quotes, books, or websites to support its claim, but the cited knowledge is incorrect and inconsistent with facts. It is essential to verify the authenticity and accuracy of the cited sources.

Figure 7: Original prompt and optimized prompt for data generation.