HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models

Junyi Li[1,3,4*], **Xiaoxue Cheng**[1*], **Wayne Xin Zhao**[1,4†], **Jian-Yun Nie**[3] and **Ji-Rong Wen**[1,2,4]

[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]School of Information, Renmin University of China
[3]DIRO, Université de Montréal
[4]Beijing Key Laboratory of Big Data Management and Analysis Methods
lijunyi@ruc.edu.cn   chengxiaoxue3@gmail.com   batmanfly@gmail.com

Abstract

Large language models (LLMs), such as Chat-GPT, are prone to generate hallucinations, *i.e.,* content that conflicts with the source or cannot be verified by the factual knowledge. To understand *what types of content* and *to which extent* LLMs are apt to hallucinate, we introduce the **Hallu**cination **Eval**uation benchmark for Large Language Models (**HaluEval**), a large collection of generated and human-annotated hallucinated samples for evaluating the performance of LLMs in recognizing hallucination. To generate these samples automatically, we propose a two-stage framework, *i.e., sampling-then-filtering*. Besides, we hire some human labelers to annotate the hallucinations in Chat-GPT responses. The empirical results suggest that ChatGPT is likely to generate hallucinated content related to specific topics by fabricating unverifiable information (*i.e.,* about 19.5% responses). Moreover, existing LLMs face great challenges in recognizing the hallucinations in texts. However, our experiments also prove that providing external knowledge or adding reasoning steps can help LLMs recognize hallucinations. Our benchmark can be accessed at https://github.com/RUCAIBox/HaluEval.

1 Introduction

The advent of large language models (LLMs) (Zhao et al., 2023) has ushered in a paradigm shift in natural language processing (NLP), making unprecedented progress in text generation and understanding (Brown et al., 2020; Li et al., 2021). The remarkable language ability makes LLMs core in a number of products with millions of users, such as the coding assistant Copilot and recent ChatGPT.

Despite these prominent capabilities of LLMs trained on large text corpus, recent work has shown that LLMs are prone to suffer from *hallucination generations* across various applications (Ji et al.,

---

User Query	Retrieve the oldest photo of a cat

ChatGPT There is no definitive answer to this question as "the oldest photo of a cat" is subjective and varies based on different sources and definitions. However, **one candidate for the oldest cat photo is the daguerreotype "Cat with Blue Ribbon," taken by photographer Antoine Claudet in 1840.** The photo depicts a cat sitting on a table, gazing directly at the camera, and wearing a blue ribbon tied around its neck. [...]

Table 1: An example from Alpaca (Taori et al., 2023) showing that ChatGPT might generate hallucinated contents (**green**) that cannot be verified by existing source.

2023; Bang et al., 2023; Sun et al., 2023), where the generated content is either in conflict with existing source or cannot be verified by the available knowledge resources. The issue of hallucination makes the deployment of LLMs potentially risky in real-world applications. Most exiting work mainly focuses on investigating the causes of hallucination for specific tasks and small language models (Cao et al., 2022; Zheng et al., 2023; Das et al., 2023). However, it still remains unclear *what types of content* and *to which extent* LLMs tend to hallucinate.

To facilitate research in this direction, we present the **Hallu**cination **Eval**uation benchmark for Large Language Models (**HaluEval**): a large collection of 35,000 hallucinated/normal samples for LLMs analysis and evaluation. HaluEval includes 5,000 *general* user queries with ChatGPT responses and 30,000 *task-specific* examples from three tasks, *i.e.,* question answering, knowledge-grounded dialogue, and text summarization. The construction pipeline of HaluEval is depicted in Figure 1. For general user queries, we adopt the 52K instruction tuning dataset from Alpaca (Taori et al., 2023) for human annotation. To further screen out user queries where LLMs are most likely to produce hallucinations, we use ChatGPT to sample three responses for each query and only retain 5,000 queries with the lowest similarity among three responses. Ac-
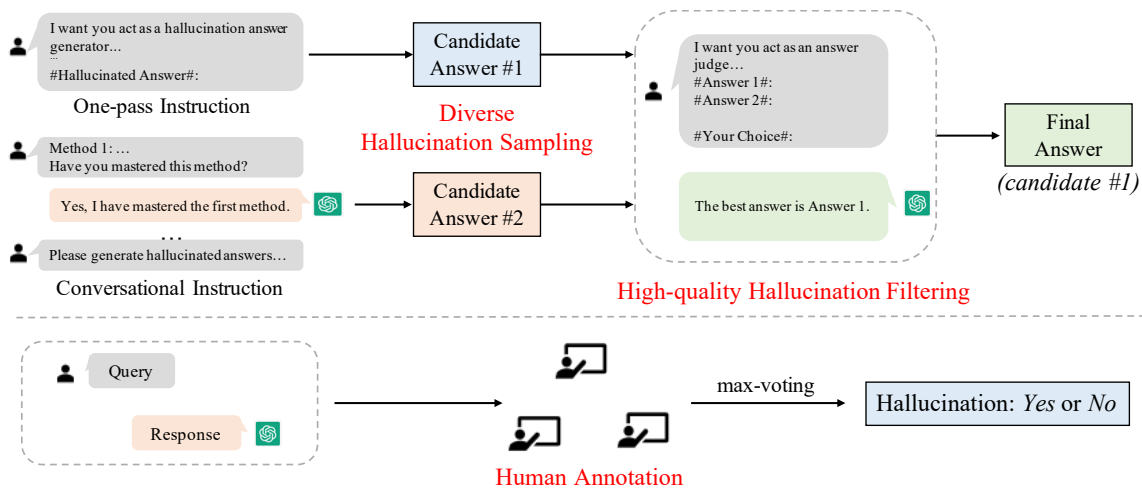
---

[*]Equal contributions
[†]Corresponding author

Figure 1: Construction pipeline of HaluEval, including automatic generation (top) and human annotation (bottom).

cording to recent work (Manakul et al., 2023), hallucinations are likely to appear in diverged and conflicting responses of LLMs. Based on the filtered user queries and ChatGPT responses, we invite human labelers to annotate whether the response contains hallucinated information and mark corresponding spans. As shown in Table 1, for the user query "*Retrieve the oldest photo of a cat*", the response generated by ChatGPT contains unverifiable information. These human-annotated queries and responses can be used to analyze what types of content LLMs tend to hallucinate and further conceive effective methods to alleviate it.

Furthermore, for the task-specific examples, we design an automatic two-stage approach to generate hallucinated samples. First, based on existing task datasets (*e.g.,* HotpotQA) as seed data, we employ ChatGPT to generate hallucinated samples with two styles of task-specific instructions, *i.e.,* one-pass and conversational. We expect that these two methods will generate diverse hallucinated samples from different aspects. Second, to select the most plausible and difficult hallucinated sample for LLMs evaluation, we elaborate the filtering instruction enhanced by ground-truth examples and leverage ChatGPT for sample selection. Through the proposed sampling-then-filtering approach, we can generate a hallucinated counterpart for each specific task example. These hallucinated samples are designed to challenge the ability of LLMs in hallucination recognition and analyze the information blind spots of LLMs.

To better understand the performance of LLMs in HaluEval, we conduct experiments with several existing powerful LLMs (*e.g.,* ChatGPT, GPT-3).

Our key findings can be summarized as follows:

• First, ChatGPT is likely to generate hallucinated content by fabricating unverifiable information in its responses (*i.e.,* about 19.5% responses). The hallucinated texts from ChatGPT cover topics including language, climate, and technology.

• Second, existing LLMs face significant challenges to identify the hallucinations in the generated text, even for ChatGPT which is used to generate these hallucinated samples (*e.g.,* only 62.59% accuracy for ChatGPT in question answering).

• Finally, the deficient performance of LLMs in recognizing hallucinations can be improved by providing explicit knowledge and adding intermediate reasoning steps. While, contrasting hallucinated samples with ground-truth makes LLMs more confused and leads to worse performance.

2 The HaluEval Benchmark

As the goal of HaluEval is to understand *what types of content* and *to which extent* LLMs tend to hallucinate, the benchmark contains a myriad of correct samples and their hallucinated counterparts. This collection is created via two ways, *i.e.,* automatic generation and human annotation.

2.1 Automatic Generation

Our generation pipeline includes two steps: 1) diverse hallucination sampling, and 2) high-quality hallucination filtering. We employ ChatGPT to execute the creation pipeline automatically.

Diverse Hallucination Sampling. Since a factual text can be hallucinated from different aspects, we propose two different hallucination sampling meth-

I want you act as a hallucination answer generator. Given a question, right answer, and related knowledge, your objective is to write a hallucinated answer that sounds plausible but is factually incorrect. You SHOULD write the hallucinated answer using the following method (each with some examples):
You are trying to answer a question but there is a factual contradiction between the answer and the knowledge. You can fabricate some information that does not exist in the provided knowledge.
#Knowledge#: The nine mile byway starts south of Morehead, Kentucky and can be accessed by U.S. Highway 60. Morehead is a home rule-class city located along US 60 (the historic Midland Trail) and Interstate 64 in Rowan County, Kentucky, in the United States. **#Question#:** What U.S Highway gives access to Zilpo Road, and is also known as Midland Trail? **#Right Answer#:** U.S. Highway 60 **#Hallucinated Answer#:** U.S. Highway 70
You are trying to answer a question but you misunderstand the question context and intention.
<Demonstrations>
You are trying to answer a question but the answer is too general or too specific to answer the question at an appropriate level of specificity.
<Demonstrations>
You are trying to answer a question but the answer cannot be inferred from the knowledge. You can incorrectly reason with the knowledge to arrive at a hallucinated answer.
<Demonstrations>
You should try your best to make the answer become hallucinated. #Hallucinated Answer# can only have about 5 more words than #Right Answer#. **#Knowledge#:** <insert the related knowledge> **#Question#:** <insert the question> **#Right Answer#:** <insert the right answer to the question> **#Hallucinated Answer#:**

Table 2: Instruction of hallucination sampling for question answering. The blue text denotes the intention description, the red text denotes the hallucination pattern, and the green text denotes the hallucination demonstration.

ods to generate diverse samples. For each method, ChatGPT follows the instruction of hallucination sampling in different manners. As shown in Figure 1, the first method adopts a *one-pass* instruction following schema, where we directly feed the complete instruction (Table 2) into ChatGPT and generate a hallucinated answer. On the other hand, the second method uses a *conversational* schema, where we teach ChatGPT to successively learn part of the instruction and make sure it has mastered. Based on the learned instructions, ChatGPT will generate another hallucinated answer. Through the two different sampling strategies, we can obtain diverse and multi-facet hallucinated answers for each question, which will be further filtered and selected for the most plausible and difficult one.

Instruction Design. In our approach, the key is to design an effective instruction for ChatGPT to generate hallucinated samples. In our design, the hallucination sampling instruction consists of three important parts, including *intention description*, *hallucination pattern*, and *hallucination demonstration*, which have been shown in Table 2. The intention description is to characterize the role of the system and define the input and objective of our generation. To control the type and quality of

hallucinated samples, we introduce the hallucination pattern and demonstration, which are related to the seed task (*e.g.,* QA in Table 2). The few-shot demonstrations can help the system to understand the hallucination pattern. In this paper, we automatically generate hallucinated samples for three tasks, *i.e.,* question answering, knowledge-grounded dialogue, and text summarization. Specifically, we consider four types of hallucination patterns for question answering (*i.e., comprehension, factualness, specificity,* and *inference*) (Zheng et al., 2023), three types of hallucination patterns for knowledge-grounded dialogue (*i.e., extrinsic-soft, extrinsic-hard,* and *extrinsic-grouped*) (Das et al., 2023), and three types of hallucination patterns for text summarization (*i.e., factual, non-factual,* and *intrinsic*) (Cao et al., 2022). For these three tasks, we first randomly sample 30,000 instances from the training set of HotpotQA (Yang et al., 2018), OpenDialKG (Moon et al., 2019), and CNN/Daily Mail (See et al., 2017), and then generate their hallucinated examples. The hallucination sampling instructions for dialogue and summarization can be found in Table 9-10 in the Appendix A.

High-quality Hallucination Filtering. To construct a challenging benchmark for LLMs, we aim

I want you act as an answer judge. Given a question, two answers, and related knowledge, your objective is to select the best and correct answer without hallucination and non-factual information. Here are some examples:

#Knowledge#: The nine mile byway starts south of Morehead, Kentucky and can be accessed by U.S. Highway 60. Morehead is a home rule-class city located along US 60 (the historic Midland Trail) and Interstate 64 in Rowan County, Kentucky, in the United States.
#Question#: What U.S Highway gives access to Zilpo Road, and is also known as Midland Trail?
#Answer 1#: U.S. Highway 60 (right answer)
#Answer 2#: U.S. Highway 70 (hallucinated answer)
#Your Choice#: The best answer is Answer 1.
...
\<Demonstrations\>

You should try your best to select the best and correct answer. If the two answers are the same, you can randomly choose one. If both answers are incorrect, choose the better one. You MUST select an answer from the provided two answers.

#Knowledge#: \<insert the related knowledge\>
#Question#: \<insert the question\>
#Answer 1#: \<insert the hallucinated answer generated by the one-pass schema\>
#Answer 2#: \<insert the hallucinated answer generated by the conversational schema\>
#Your Choice#:

Table 3: Instruction of hallucination filtering for question answering.

to select the most plausible and difficult halluci-nated samples from the above two sampling meth-ods. As shown in Table 3, we design the instruction of hallucination filtering enhanced by ground-truth answers to select the best answer from two hallu-cinated candidates. In the instruction of filtering, the demonstration includes the ground-truth correct answer (*e.g.,* U.S. Highway 60) and a hallucinated counterpart (*e.g.,* U.S. Highway 70). While, in the test example, we input two hallucinated answers. Following the demonstrations, we expect ChatGPT to select one of the hallucinated answers that is the most plausible and closest to the right answer. Fi-nally, the selected hallucinated sample is hard to be identified, which are further used to evaluate LLMs in hallucination recognition. The instructions of hallucination filtering for dialogue and summariza-tion are shown in Table 11-12 in the Appendix B.

Through the *sampling-then-filtering* process, we end up generating a total of $30,000$ hallucinated samples for the three tasks. Our approach can also be adapted to other tasks and datasets.

2.2 Human Annotation

Besides generating hallucinated samples, we also invite human labelers to annotate whether ChatGPT responses contain hallucinated content.

We annotate the general user queries and Chat-GPT responses from the $52K$ instruction tuning dataset from Alpaca (Taori et al., 2023), which has been widely used by recent LLMs. To screen out user queries where LLMs are most likely to produce hallucination for labeling, we design a pre-

Question	In what political party was the man who officially opened Royal Spa Centre in 1972?
Right Answer	Conservative
Hallucinated Answer	Labour Party
User Query	Retrieve the oldest photo of a cat
ChatGPT	There is no definitive answer to this question as "the oldest photo of a cat" is subjective and varies based on different sources and definitions. However, one candidate for the oldest cat photo is the daguerreotype "Cat with Blue Ribbon," taken by photographer Antoine Claudet in 1840. The photo depicts a cat sitting on a table, gazing directly at the camera, and wearing a blue ribbon tied around its neck. [...]
Hallucination	Yes
Fragments	the oldest cat photo is the daguerreotype "Cat with Blue Ribbon" taken by photographer Antoine Claudet in 1840.

Table 4: A generated hallucinated QA example and a human-labeled ChatGPT response for a user query.

selection procedure. Specifically, we use ChatGPT to sample three responses for each user query and compute their average semantic similarity using BERTScore (Zhang et al., 2020). We finally re-tain $5,000$ user queries with the lowest similarities. According to recent work (Manakul et al., 2023), hallucinations are likely to appear in diverged and conflicting responses of LLMs. For each query and ChatGPT response, human labelers will annotate whether the response contains hallucinated infor-mation ("Yes" or "No") and list the corresponding

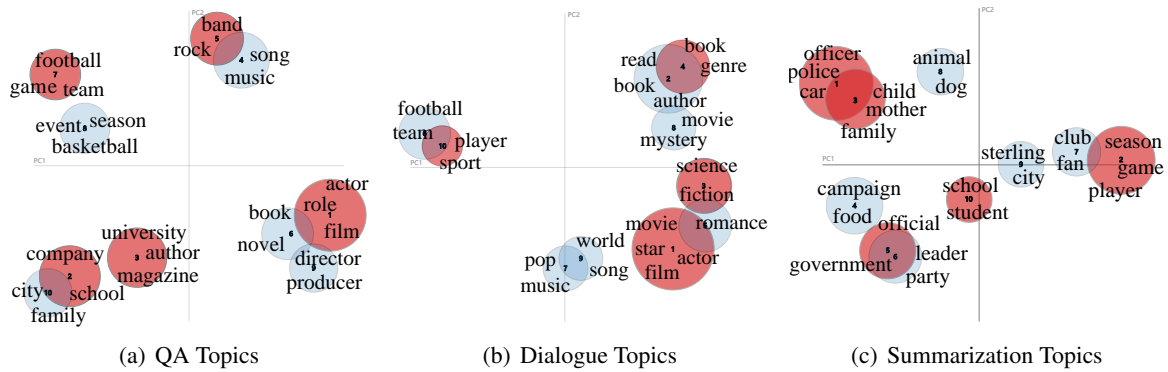(a) QA Topics   (b) Dialogue Topics   (c) Summarization Topics

Figure 2: Topic distributions for QA, knowledge-grounded dialogue, and text summarization. The samples of each task are classified into 10 topics, and the red circles denote the topics of failed recognized samples by ChatGPT.
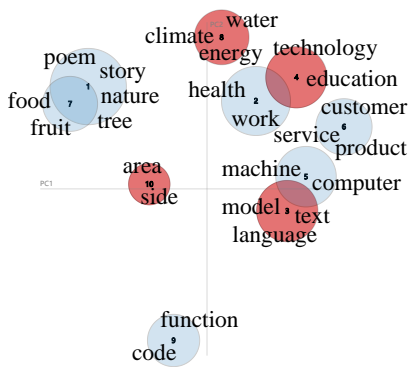

Figure 3: Topic distribution for ChatGPT responses.

spans. The hallucination is considered from the following three aspects: *unverifiable*, *non-factual*, and *irrelevant*. Each response is labeled by three human labelers, and we adopt the max-voting strategy to determine the final hallucination label.

Labeler Details. Annotating the hallucination in ChatGPT responses is a very challenging task, which requires good reading comprehension skills and using search engine to look up relevant information for judgement. Thus, from an initial pool of labeler candidates, we select labelers who are good at English passage reading with at least an undergraduate-level education. Besides, following (Ouyang et al., 2022), we have labelers annotate a small number of test examples and measure their agreement with the labels of researchers, and finally we choose thirty human labelers with the highest agreement scores. We report Fleiss's Kappa (κ) to indicate the reliability of agreement between human labelers. We compute κ on $5,000$ annotated samples and obtain $\kappa = 0.811$ ($0.80 \leq \kappa \leq 1.00$) showing a perfect agreement.

2.3 Benchmark Analysis and Usage

With the automatic two-step generation process in Section 2.1, we produce a total of $30,000$ hallucinated samples with $10,000$ examples for each task of QA, dialogue, and summarization. We show the number of generated samples for each hallucination pattern in Table 16 at the Appendix D. Moreover, we manually annotate $5,000$ ChatGPT responses for general user queries in Section 2.2. We present a QA example and an annotated query and response example in Table 4. Among the annotated ChatGPT responses, 977 responses are labeled as containing hallucination (19.5%). Finally, we present the topic distributions of our generated task-specific samples and annotated ChatGPT responses in Figure 2 and Figure 3, ranging from film, sports to school, computer, technology, etc.

With our benchmark, researchers can use it to investigate or mitigate the hallucination issue for LLMs in three aspects. Firstly, based on our generated and annotated samples, researchers can use them to analyze what types of content LLMs tend to generate hallucinations. Second, researchers can further evaluate the ability of LLMs to recognize hallucinations in the generated samples. For example, given a question and an answer, LLMs can be asked to determine whether the answer contains hallucinated content. Finally, our benchmark can be further paired with human annotation to assess whether the LLMs' output contains hallucinations, since the samples in our benchmark are specially designed for testing the hallucinations of LLMs.

To use our benchmark, users can run the code in our project repository to conduct the corresponding evaluation and analysis. Users can use our provided instructions on their own datasets to evaluate LLMs on hallucinations.

Models	QA	Dialogue	Summarization	General
ChatGPT	62.59	72.40	58.53	79.44
Claude 2	69.78	64.73	57.75	75.00
Claude	67.60	64.83	53.76	73.88
Davinci002	60.05	60.81	47.77	80.42
Davinci003	49.65	68.37	48.07	80.40
GPT-3	49.21	50.02	51.23	72.72
Llama 2	49.60	43.99	49.55	20.46
ChatGLM	47.93	44.41	48.57	30.92
Falcon	39.66	29.08	42.71	18.98
Vicuna	60.34	46.35	45.62	19.48
Alpaca	6.68	17.55	20.63	9.54

Table 5: Accuracy (%) of classifying whether a sample contains hallucinated contents.

3 Experiments

3.1 Experimental Setup

Evaluation Models. We evaluate several state-of-the-art LLMs in HaluEval benchmark. First, we experiment on five closed-source LLMs, including OpenAI's GPT-3 (davinci) (Brown et al., 2020), InstructGPT (text-davinci-002/003) (Ouyang et al., 2022), ChatGPT (gpt-3.5-turbo) and Anthropic's Claude and Claude 2 models, which can only be accessed through their APIs. Besides, we also evaluate five prevalent open-source LLMs, including Alpaca (7B) (Taori et al., 2023), Vicuna (7B) (Chiang et al., 2023), ChatGLM (7B) (Zeng et al., 2022), Falcon (7B) (TII, 2023), and Llama 2-Chat (7B) (Touvron et al., 2023). Our experiments were performed without fine-tuning or engaging in the tuning of hyper-parameters.

Implementation Details. We execute the generation process of hallucinated samples using Azure OpenAI ChatGPT API. We use a temperature of 1.0 to generate samples and set the maximum number of tokens for generation to 256. Moreover, we set the frequency penalty to zero and top-p to 1.0. For evaluation, we set the temperature to zero for all models to reduce output randomness and ensure more focused and deterministic outputs.

In the following, we first conduct hallucination recognition experiments, then propose several potentially useful strategies to improve the recognition, and finally we perform qualitative analysis to understand the hallucination in LLMs.

3.2 Results and Analysis

3.2.1 Hallucination Recognition

To evaluate the ability of LLMs to recognize hallucinations, we randomly select the hallucinated

or normal output (*e.g.,* an answer) of each sample for classification. The evaluation instructions of QA, dialogue, and summarization are presented in Table 13, Table 14 and Table 15 in Appendix C.

Table 5 presents the accuracy of evaluated LLMs to classify whether the sample output contains hallucinated information. Our findings indicate that LLMs are still poor at identifying hallucination which might be implicit in text. For example, the state-of-the-art ChatGPT model cannot distinguish between factual and hallucinated summary and only achieves 58.53% accuracy in text summarization, which is barely above chance. Moreover, GPT-3 obtains just about random chance of 50% accuracy across three tasks, and Alpaca or Vicuna even performs worse (well below random chance). We hypothesize that LLMs perform poorly because the hallucinated sample we generate looks highly similar with ground-truth ones but differs in the key factual spans. As we can see, from GPT-3 to InstructGPT and ChatGPT, instruction tuning and alignment with humans can strength the ability of LLMs in identifying the hallucinations in text.

With respect to the hallucinated samples where ChatGPT fails to recognize, we present the number of each hallucination pattern in Table 6. Based on the results, we can observe that the hallucination patterns of failed samples are unevenly distributed. For example, over half of failures in QA, dialogue, and summarization originate from the first hallucination pattern (*i.e., comprehension, extrinsic-soft,* and *factual*), which refers to the hallucinations that are factually correct but conflict with the context. This indicates that LLMs lack or cannot associate related knowledge to identify the factual hallucination in the generated text. To further understand the failures of ChatGPT, we visualize the topics of those failed samples via Latent Dirichlet Allocation (LDA) (Blei et al., 2003). As shown in Figure 2 and Figure 3, we cluster all task samples into ten topics and mark the topics of failed samples as red. We find that the hallucination of LLMs is topic-sensitive. For example, the frequent topics in QA include film, school, and company. While, ChatGPT mainly fails to recognize those samples in the topics of film, company, and band. For user queries and ChatGPT responses, the top five topics include story, health, language, technology, and computer. ChatGPT mainly faces challenges in topics of technology, climate, and language.

Tasks	#Failed	P-I	P-II	P-III	P-IV
QA	3109	1559	245	278	1027
Dialogue	891	465	344	82	-
Summarization	3868	3106	705	57	-

Table 6: Number of samples where ChatGPT fails to recognize for each hallucination pattern (P-I/II/III/IV).

3.2.2 Improvement Strategies

In this part, we design several strategies to improve the ability of LLMs to recognize hallucination. The results are shown in Table 8.

Knowledge Retrieval. Retrieving relevant knowledge is a widely used strategy to eliminate hallucination (Lewis et al., 2020; Li et al., 2023a). Therefore, we supply ChatGPT with the knowledge facts retrieved from Wikipedia (except for that summarization does not need external information besides the source document). By providing knowledge, the recognition accuracy of ChatGPT increases significantly (e.g., increasing from 62.59 to 76.83 in QA), while the performance improvement in dialogue is mild. We hypothesize that the common hallucination patterns in dialogue (i.e., extrinsic-soft/hard) cannot be simply identified via incorporating external knowledge. For those general user queries and ChatGPT responses, we discover that providing external knowledge does have a significant benefit. Thus, equipping LLMs with external knowledge can largely enhance their abilities to recognize hallucinations.

CoT Reasoning. In previous work (Wei et al., 2022), chain-of-thought (CoT) has been proposed to improve the ability of LLMs to perform reasoning and derive the final answer by introducing a series of intermediate reasoning steps. Here, besides producing the recognition result, we also require ChatGPT to generate the reasoning steps. While, from the results in Table 8, we observe that generating reasoning steps can mildly improve the performance but makes the model perform worse in QA and dialogue (e.g., dropping from 62.59 to 59.58). Compared to retrieving knowledge, adding chain-of-thought before output might interfere with the final judgement. While, in text summarization, generating reasoning steps improve the accuracy from 58.53 to 61.21. The reason might be that the factual contradiction between document and summary can be identified through logic reasoning.

Sample Contrast. We further provide ground-truth examples for ChatGPT to test whether it can distinguish the right sample from the hallucinated sample. As we can see from Table 8, distinguishing between right and hallucinated samples achieves the worst results. We hypothesize that our generated hallucinated samples have a high similarity to the real samples, thus making LLMs confused to distinguish them. This test also indicates that our benchmark is very challenging in hallucination evaluation for LLMs.

3.3 Case Study

In the above, we have observed that providing external knowledge can be beneficial for LLMs to mitigate and recognize hallucinations. To demonstrate the effectiveness of knowledge retrieval in mitigating hallucinations, we present two hallucinated responses from ChatGPT and refined responses after augmented with retrieved knowledge in Table 7. In the first example, the generated span (i.e., "July 4, 1776 - Declaration of Independence signing") contains hallucinated information because it gives a wrong time of Declaration of Independence signing. By providing retrieved information about Declaration of Independence signing, ChatGPT is able to correct the hallucinated span and give the right information. Analogously, in the second example, ChatGPT gives incorrect GDP growth rates of China and India, which is due to that API-based ChatGPT cannot access the web to obtain the official data. After providing official information retrieved from World Bank, the refined span displays answers that contain the correct information. The above two examples illustrate that retrieving knowledge related to queries can help ChatGPT significantly reduce the hallucinations in the response, especially those factual errors.

4 Related Work

Hallucination in LLMs. Hallucination in LLMs is concerning since it hinders performance and raises safety risks in real-world application. To alleviate this issue, prior studies have proposed to use a verification system to identify non-factual entities in text summarization (Zhao et al., 2020), invoke interfaces of structured data (e.g., knowledge graph, database) to obtain related evidence (Jiang et al., 2023; Lan et al., 2022), and train a token-level fact critic to recognize hallucination and rectify them in dialogue (Dziri et al., 2021). To enhance the understanding of hallucination in LLMs and pro-

User Query	Generate a list of 5 important dates in US history.	Create a visualization to compare the GDP growth of India and China between 1998 and 1998.
ChatGPT Response	**1. July 4, 1776 - Declaration of Independence signing** 2. April 12, 1861 - Start of the Civil War 3. August 18, 1920 - 19th Amendment ratification granting women the right to vote 4. December 7, 1941 - Pearl Harbor attack during World War II 5. September 11, 2001 - Terrorist attacks on New York City and Washington D.C.	As an AI language model, I cannot create visualizations, but I can provide you with the following information: According to World Bank data, India and China have shown significant economic growth in the past two decades. **In 1998, India's GDP growth rate was 4.8%, while China's GDP growth rate was 9.3%.** In the following years, the GDP growth rate fluctuated, but China maintained a higher growth rate than India. **By 2019, China's GDP growth rate was 6.1%, while India's GDP growth rate was 4.2%.**
Retrieved Knowledge	The signing of the United States Declaration of Independence occurred primarily on August 2, 1776, at the Pennsylvania State House in Philadelphia, later known as Independence Hall. https://en.wikipedia.org/wiki/Signing_of_the_United_States_Declaration_of_Independence	India's GDP growth rate in 1998 was 6.2%. China's GDP growth rate in 1998 was 7.8%. India's GDP growth rate in 2019 was 3.7%. China's GDP growth rate in 2019 was 6.0%. https://data.worldbank.org/
Refined Response	**1. August 2, 1776 - Declaration of Independence signing** 2. April 12, 1861 - Start of the Civil War 3. August 18, 1920 - 19th Amendment ratification granting women the right to vote 4. December 7, 1941 - Pearl Harbor attack during World War II 5. September 11, 2001 - Terrorist attacks on New York City and Washington D.C.	According to World Bank data, India and China have shown significant economic growth in the past two decades. **In 1998, India's GDP growth rate was 6.2%, while China's GDP growth rate was 7.8%. By 2019, China's GDP growth rate was 6.0%, while India's GDP growth rate was 3.7%.**

Table 7: Two hallucinated and refined examples from ChatGPT. The **green** text denotes the hallucinated span, and the **brown** text denotes the refined span after augmented with retrieved knowledge.

Variants	QA	Dialogue	Summarization	General
ChatGPT	62.59	72.40	58.53	86.22
w/ Knowledge	76.83	73.80	-	90.73
w/ CoT	59.58	71.39	61.21	86.50
w/ Contrast	49.19	68.67	49.46	-

Table 8: Accuracy (%) of ChatGPT equipped with three improvement strategies.

mote the unification of research efforts, there are many active endeavors to analyze the causes of hallucination in different tasks and investigate their relationship (Zheng et al., 2023; Das et al., 2023; Cao et al., 2022). Our work is closely related to these work, but we focus on building a hallucination evaluation benchmark for LLMs. Our dataset can serve as a public platform for exhibiting the blind spots of LLMs in solving hallucination.

Hallucination Evaluation. Another line of work focusing on evaluating the hallucination of models in different NLP tasks (Dziri et al., 2022b; Gupta et al., 2022; Dziri et al., 2022a; Rashkin et al., 2021; Li et al., 2023b). For instance, The BEGIN benchmark (Dziri et al., 2022b) classifies the utterances generated by dialogue systems into three categories,

i.e., fully attributable, not fully attributable, and generic; and the Attributable to Identified Sources (AIS) benchmark (Rashkin et al., 2021) assesses whether the source documents support the output of text generation models. Though these benchmarks can serve as decent evaluation platform, they are penurious in only focusing on single tasks (*e.g.,* dialogue) and small models (*e.g.,* DPR). Besides, several metrics have been proposed to quantify hallucination, such as PARENT (Dhingra et al., 2019) for measuring n-gram lexical entailment in table-to-text generation and TRUE (Honovich et al., 2022) computes the example-level Area Under the ROC Curve. In this work, our HaluEval benchmark includes general user queries and ChatGPT responses and proposes a two-step automatic process to generate hallucinated samples for evaluation, which is completely based on LLMs.

5 Conclusion

We introduce HaluEval, a large-scale collection of generated and human-annotated hallucinated samples for evaluating the performance of LLMs in recognizing hallucinations. To automatically generate large-scale samples, we propose a two-step

approach, *i.e.,* sampling-then-filtering. We first introduce two different sampling methods to generate diverse samples using instructions and then filter and select the difficult one. Besides, we invite qualified human labelers to annotate the hallucinations of ChatGPT responses given user queries. We find that, existing LLMs mostly fail to recognize the hallucinations in text and tend to generate hallucinated content. Finally, we suggest several strategies to help LLMs recognize hallucinations. Our benchmark can facilitate research in understanding what types of content and to which extent LLMs tend to hallucinate, ultimately paving the way for building more effective and reliable LLMs in the future.

6 Limitations

In our approach, we leverage a LLM, *i.e.,* ChatGPT, to automatically generate the hallucinated samples. Therefore, the quality of our hallucinated samples is limited by the capacity of ChatGPT in following the complex instruction of hallucination sampling. Although we design the high-quality hallucination filtering process, it is still necessary to apply quality control to the generation of hallucinated samples. Besides, our benchmark focuses on evaluating the ability of LLMs in recognizing the hallucinations in text but does not investigate the underlying reasons behind the appearance of hallucinations like prior work (Zheng et al., 2023; Das et al., 2023).

As for the potential issue, since the hallucinated samples in our benchmark looks highly similar to the ground-truth samples, which might be misused for an unexpected purpose than we planned. To alleviate this issue, we should monitor and regulate the spread and usage of our benchmark.

Acknowledgments

References

2023. Introducing Falcon LLM . https://falconllm.tii.ae.

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. *CoRR*, abs/2302.04023.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354.

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 opensource chatbot impressing gpt-4 with 90%* chatgpt quality.

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

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4884–4895. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar R. Zaïane, Mo Yu, Edoardo Maria Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Trans. Assoc. Comput. Linguistics*, 10:1473–1490.

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.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Trans. Assoc. Comput. Linguistics*, 10:1066–1083.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3785–3801. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3905–3920. 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 Computing Surveys*, 55(12):1–38.

Jinhao Jiang, Kun Zhou, Keming Ye Zican Dong, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason on structured data.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–20.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

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

Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Pretrained language model for text generation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4492–4499. ijcai.org.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *CoRR*, abs/2303.08896.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Long Ouyang, Jeffrey 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. In *NeurIPS*.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. Measuring attribution in natural language generation models. *CoRR*, abs/2112.12870.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13618–13626. AAAI Press.

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.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: an open bilingual pre-trained model. abs/2210.02414.

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.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *CoRR*, abs/2009.13312.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in answering questions faithfully? *CoRR*, abs/2304.10513.

Appendix

We provide some extra information about our benchmark as supplementary materials. The appendix is organized into three sections:

- Instructions of hallucination sampling are presented in Appendix A;

- Instructions of hallucination filtering are presented in Appendix B;

- Instructions of evaluation are presented in Appendix C;

- Details of our benchmark are presented in Appendix D.

A Hallucination Sampling

The hallucination sampling instructions for dialogue and summarization are shown in Table 9 and Table 10, respectively.

B Hallucination Filtering

The hallucination sampling instructions for dialogue and summarization are shown in Table 11 and Table 12, respectively.

C Hallucination Recognition

The hallucination recognition instructions for QA, dialogue and summarization are shown in Table 13, Table 14 and Table 15, respectively.

D Details of HaluEval

The number of generated hallucinated samples for each hallucination pattern are shown in Table 16.

I want you act as an assistant in a conversation with human. Given a dialogue history, the true response, and related knowledge, your objective is to write a hallucinated response that sounds plausible but is factually incorrect. You SHOULD write the hallucinated response using the following method (each with some examples):

You are trying to write a response to human but you replace the true entity with a highly similar entity.

#Knowledge#: The Dark Knight is a 2008 superhero film directed by Christopher Nolan from a screenplay he co-wrote with his brother Jonathan. Christopher Nolan is a film director.
#Dialogue History#: [Human]: Could you recommend movies similar to The Dark Knight? [Assistant]: The sequel to Batman Begins is The Dark Knight. [Human]: Okay. Who is the director of The Dark Knight and any other movies from him not related to Batman?
#True Response#: Christopher Nolan was the director. He also directed insomnia and inception.
#Hallucinated Response#: Steven Spielberg was the director. He also directed insomnia and inception.

or

You are trying to write a response to human but you replace the true entity with a dissimilar entity.

<Demonstrations>

or

You are trying to write a response to human but you replace the true entity with a dissimilar entity in a different entity type.

<Demonstrations>

You should try your best to make the response become hallucinated.

#Knowledge#: <Here is the related knowledge>
#Dialogue History#: <Here is the dialogue history>
#True Response#: <Here is the true response of the dialogue history>
#Hallucinated Response#:

Table 9: Instruction of hallucination sampling for knowledge-grounded dialogue.

I want you act as a hallucination summary generator. Given a document and the right summary, your objective is to write a hallucinated summary that sounds plausible but is factually incorrect. You SHOULD write the hallucinated summary using the following method (each with some examples):

You are trying to write a summary which is factual but some information cannot be directly inferred or entailed from the document.

#Document#: The panther chameleon was found on Monday by a dog walker in the wooded area at Marl Park. It had to be put down after X-rays showed all of its legs were broken and it had a deformed spine. RSPCA Cymru said it was an "extremely sad example of an abandoned and neglected exotic pet". Inspector Selina Chan said: "It is a possibility that the owners took on this animal but were unable to provide the care he needs and decided to release him to the wild. "We are urging potential owners of exotic animals to thoroughly research what is required in the care of the particular species before taking one on. "Potential owners need to make sure they can give their animal the environment it needs and they have the facilities, time, financial means and long-term commitment to maintain a good standard of care, as required under the Animal Welfare Act 2006." She added it was illegal to release non-native species into the wild.
#Right Summary#: Owners of exotic animals have been urged to do research before having them as pets after a seriously neglected chameleon was found in Cardiff Bay.
#Hallucinated Summary#: A chameleon that was found in a Cardiff park has been put down after being abandoned and neglected by its owners.

or

You are trying to write a summary but there exist some non-factual and incorrect information. You can fabricate some information that does not exist in the provided document.

<Demonstrations>

or

You are trying to write a summary but there is a factual contradiction between the summary and the document.

<Demonstrations>

You should try your best to make the summary become hallucinated. #Hallucinated Summary# can only have about 5 more words than #Right Summary#.

#Document#: <Here is the test document>
#Right Summary#: <Here is the right summary of the test document>
#Hallucinated Summary#:

Table 10: Instruction of hallucination sampling for text summarization.

I want you act as a response judge. Given a dialogue history, two responses, and related knowledge, your objective is to select the best and correct response without hallucination and non-factual information. Here are some examples:

#Knowledge#: The Dark Knight is a 2008 superhero film directed by Christopher Nolan from a screenplay he co-wrote with his brother Jonathan. Christopher Nolan is a film director.
#Dialogue History#: [Human]: Could you recommand movies similar to The Dark Knight? [Assistant]: The sequel to Batman Begins is The Dark Knight. [Human]: Okay. Who is the director of The Dark Knight and any other movies from him not related to Batman?
#Response 1#: Christopher Nolan was the director. He also directed insomnia and inception.
#Response 2#: Steven Spielberg was the director. He also directed insomnia and inception.
#Your Choice#: The best response is Response 1.
...
<Demonstrations>
...

You should try your best to select the best and correct response. If the two responses are the same, you can randomly choose one. If both responses are incorrect, choose the better one. You MUST select a response from the provided two responses.

#Knowledge#: <Here is the related knowledge>
#Dialogue History#: <Here is the dialogue history>
#Response 1#: <Here is the hallucinated response generated by the first channel>
#Response 2#: <Here is the hallucinated response generated by the second channel>
#Your Choice#:

Table 11: Instruction of hallucination filtering for knowledge-grounded dialogue.

I want you act as a summary judge. Given a document and two summaries, your objective is to select the best and correct summary without hallucination and non-factual information. Here are some examples:

#Document#: The panther chameleon was found on Monday by a dog walker in the wooded area at Marl Park. It had to be put down after X-rays showed all of its legs were broken and it had a deformed spine. RSPCA Cymru said it was an "extremely sad example of an abandoned and neglected exotic pet". Inspector Selina Chan said: "It is a possibility that the owners took on this animal but were unable to provide the care he needs and decided to release him to the wild. "We are urging potential owners of exotic animals to thoroughly research what is required in the care of the particular species before taking one on. "Potential owners need to make sure they can give their animal the environment it needs and they have the facilities, time, financial means and long-term commitment to maintain a good standard of care, as required under the Animal Welfare Act 2006." She added it was illegal to release non-native species into the wild.
#Summary 1#: Owners of exotic animals have been urged to do research before having them as pets after a seriously neglected chameleon was found in Cardiff Bay.
#Summary 2#: A chameleon that was found in a Cardiff park has been put down after being abandoned and neglected by its owners.
#Your Choice#: The best summary is Summary 1.
...
<Demonstrations>
...

You should try your best to select the best and correct summary. If both summaries are incorrect, choose the better one. You MUST select a summary from the provided two summaries.

#Document#: <Here is the test document>
#Summary 1#: <Here is the hallucinated summary generated by the first channel>
#Summary 2#: <Here is the hallucinated summary generated by the second channel>
#Your Choice#:

Table 12: Instruction of hallucination filtering for text summarization.

I want you act as an answer judge. Given a question and an answer, your objective is to determine if the provided answer contains non-factual or hallucinated information. You SHOULD give your judgement based on the following hallucination types and the world knowledge.

You are trying to determine if there is a factual contradiction between the answer and the world knowledge. Some information in the answer might be fabricated.

#Question#: What U.S Highway gives access to Zilpo Road, and is also known as Midland Trail?
#Answer#: U.S. Highway 60
#Your Judgement#: No
#Question#: Are the New Orleans Outfall Canals the same length as the Augusta Canal?
#Answer#: No. The Orleans Canal is approximately 3.6 miles (5.8 kilometers) long while the Augusta Canal is approximately 7 miles (11.3 kilometers) long.
#Your Judgement#: Yes

You are trying to determine if the answer misunderstands the question context and intention.
<Demonstrations>

You are trying to determine if the answer is too general or too specific to answer the question at an appropriate level of specificity.
<Demonstrations>

You are trying to determine if the answer cannot be inferred from the knowledge correctly.
<Demonstrations>

You should try your best to determine if the answer contains non-factual or hallucinated information according to the above hallucination types. The answer you give MUST be "Yes" or "No".

#Question#: <Here is the test question>
#Answer#: <Here is the hallucinated answer or right answer>
#Your Judgement#:

Table 13: Instruction of hallucination recognition for question answering.

I want you act as a response judge. Given a dialogue history and a response, your objective is to determine if the provided response contains non-factual or hallucinated information. You SHOULD give your judgement based on the following hallucination types and the world knowledge.

You are trying to determine if the true entity in the response is replaced with a highly similar entity.
#Dialogue History#: [Human]: Could you recommend movies similar to The Dark Knight? [Assistant]: The sequel to Batman Begins is The Dark Knight. [Human]: Okay. Who is the director of The Dark Knight and any other movies from him not related to Batman?
#Response#: Christopher Nolan was the director. He also directed insomnia and inception.
#Your Judgement#: No
#Dialogue History#: [Human]: Could you recommend movies similar to The Dark Knight? [Assistant]: The sequel to Batman Begins is The Dark Knight. [Human]: Okay. Who is the director of The Dark Knight and any other movies from him not related to Batman?
#Response#: Steven Spielberg was the director. He also directed insomnia and inception.
#Your Judgement#: Yes

You are trying to determine if the true entity in the response is replaced with a dissimilar entity.
<Demonstrations>

You are trying to determine if the true entity in the response is replaced with a dissimilar entity in a different entity type.
<Demonstrations>

You should try your best to determine if the response contains non-factual or hallucinated information according to the above hallucination types. The answer you give MUST be "Yes" or "No".

#Dialogue History#: <Here is the dialogue history>
#Response#: <Here is the hallucinated response or right response>
#Your Judgement#:

Table 14: Instruction of hallucination recognition for knowledge-grounded dialogue.

I want you act as a summary judge. Given a document and a summary, your objective is to determine if the provided summary contains non-factual or hallucinated information. You SHOULD give your judgement based on the following hallucination types and the world knowledge.

You are trying to determine if the summary is factual but some information cannot be directly inferred or entailed from the document.

#Document#: The panther chameleon was found on Monday by a dog walker in the wooded area at Marl Park. It had to be put down after X-rays showed all of its legs were broken and it had a deformed spine. RSPCA Cymru said it was an "extremely sad example of an abandoned and neglected exotic pet". Inspector Selina Chan said: "It is a possibility that the owners took on this animal but were unable to provide the care he needs and decided to release him to the wild. "We are urging potential owners of exotic animals to thoroughly research what is required in the care of the particular species before taking one on. "Potential owners need to make sure they can give their animal the environment it needs and they have the facilities, time, financial means and long-term commitment to maintain a good standard of care, as required under the Animal Welfare Act 2006." She added it was illegal to release non-native species into the wild.
#Summary#: A chameleon that was found in a Cardiff park has been put down after being abandoned and neglected by its owners.
#Your Judgement#: Yes

You are trying to determine if there exists some non-factual and incorrect information in the summary.
<Demonstrations>

You are trying to determine if there is a factual contradiction between the summary and the document.
<Demonstrations>

You should try your best to determine if the summary contains non-factual or hallucinated information according to the above hallucination types. The answer you give MUST be "Yes" or "No".

#Document#: <Here is the test document>
#Summary#: <Here is the hallucinated summary or right summary>
#Your Judgement#:

Table 15: Instruction of hallucination recognition for text summarization.

Tasks	#Sample	P-I	P-II	P-III	P-IV
QA	10000	2280	1378	5102	1240
Dialogue	10000	8330	1196	474	-
Summa.	10000	2614	3562	3824	-

Table 16: Number of generated samples for each hallucination pattern (P-I/II/III/IV). "'Summa." is short for summarization. "-" is due to that we consider three patterns in dialogue and summarization.