DiaHalu: A Dialogue-level Hallucination Evaluation Benchmark for Large Language Models

Anonymous ACL submission

Abstract

Since large language models (LLMs) achieve significant success in recent years, the hallucination issue remains a challenge, numerous benchmarks are proposed to detect the hallucination. Nevertheless, some of these benchmarks are not naturally generated by LLMs but are intentionally induced. Also, many merely focus on the factuality hallucination while ignoring the faithfulness hallucination. Additionally, although dialogue pattern is more widely utilized in the era of LLMs, current benchmarks only concentrate on sentence-level and passage-level hallucination. In this study, we propose DiaHalu, the first dialogue-level hallucination evaluation benchmark to our knowledge. Initially, we integrate the collected topics into system prompts and facilitate a dialogue between two ChatGPT3.5. Subsequently, we manually modify the contents that do not adhere to human language conventions and then have LLMs re-generate, simulating authentic human-machine interaction scenarios. Finally, professional scholars annotate all the samples in the dataset. DiaHalu covers four common multi-turn dialogue domains and five hallucination subtypes, extended from factuality and faithfulness hallucination. Experiments through some well-known LLMs and detection methods on the dataset show that DiaHalu is a challenging benchmark, holding significant value for further research.

1 Introduction

Large language models (LLMs) (Zhao et al., 2023a), addressing many downstream tasks through natural language generation (NLG) technique, achieve significant success across diverse fields of natural language processing (NLP) (Li et al., 2024; Pan et al., 2024; Wei et al., 2023; Fan et al., 2023). With a considerable volume of parameters and sophisticated training methodologies, LLMs significantly propelled advancements in artificial intelligence (Zhao et al., 2023a).

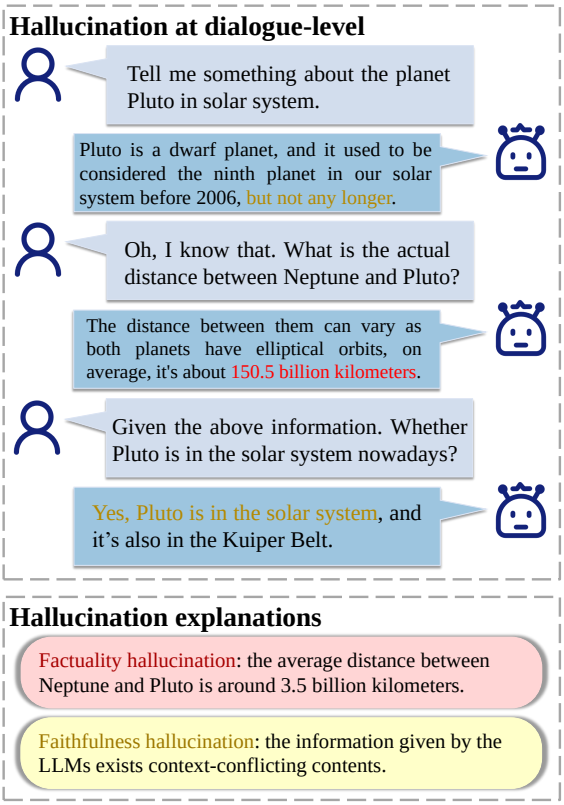

Figure 1: Our benchmark not only includes factuality hallucination but also incorporates faithfulness hallucination at the dialogue level, although most benchmarks overlook the latter.

Despite many advantages of large language models, the issue of hallucination remains a primary concern (Ji et al., 2023; Zhang et al., 2023). Hallucination predominantly centers on the inclination of LLMs to generate nonsensical or untruthful contents for specific sources (Wang et al., 2023). Therefore, the occurrence of hallucinations poses risks to the application of large language models in various real-world scenarios (Omiye et al., 2023; Kasneci et al., 2023; Wu et al., 2023a).

Given the aforementioned risks, hallucination detection emerges as a highly crucial task. In recent

years, researchers propose numerous benchmarks for hallucination detection task (Li et al., 2023; Guan et al., 2023; Manakul et al., 2023; Yang et al., 2023). Nevertheless, several problems persist in these benchmarks. **(1) Not naturally generated.** One pitfall of existing benchmarks for detecting LLMs' hallucination is that the hallucinated contents are typically induced via manually designed trigger prompts (Li et al., 2023), while not naturally generated by LLMs as in daily usage (Liu et al., 2022). **(2) Merely focusing on factuality hallucination.** Most benchmarks merely focus on detecting factuality hallucination (Guan et al., 2023), with few datasets that can demonstrate faithfulness hallucination (Huang et al., 2023a) (Figure 1). **(3) Only concentrating on sentence-level and passage-level.** Researchers proposes many sentence-level (Manakul et al., 2023; Zhao et al., 2023b) and passage-level (Yang et al., 2023; Feng et al., 2023b) hallucination detection benchmarks. However, the dialogue pattern has broader and more widespread applications in LLMs. Although it is not mentioned in previous researches to our current knowledge, dialogue-level hallucination detection is equally essential.

Therefore, we propose a new **dia**logue-level **hallu**cination evaluation benchmark for large language models (**DiaHalu**). We **initially** determine four domains for multi-turn dialogue: knowledge-grounded, task-oriented, chit-chat and reasoning. For each domain, we undertake a three-step process to construct the dataset. (1) We collect topics for dialogues from various sources. (2) We incorporate the topics into artificially designed system prompts and input them into two ChatGPT3.5, enabling them to engage in a multi-turn dialogue. (3) Since the knowledge-grounded and task-oriented domains stand for real human-machine interaction scenarios, we align the contents of one of the conversational participants with human language. We manually modify the contents that do not conform to human language conventions and have LLMs re-generate, resulting in the final outcomes. **Then**, professional scholars annotate all the samples with labels, hallucination types and locations, as well as explanations. It is noteworthy that we not only consider the factuality hallucination but also further classify the faithfulness hallucination into three types: Incoherence, Irrelevance and Overreliance. We similarly introduce the reasoning hallucination for the reasoning domain. **Additionally**, we conduct experiments on DiaHalu by deploying existing hallucination detection methods and some commonly used LLMs. The results indicate that DiaHalu is a highly challenging dataset, and both Chain-of-Thought and retrieval methods can improve the performance of the detection. Our contributions can be listed as follows:

- To our current knowledge, we propose the first dialogue-level hallucination detection benchmark named DiaHalu[1].
- DiaHalu covers four multi-turn dialogue domains along with five hallucination subtypes extended from factuality and faithfulness hallucinations, which are more widely applicable in real-world scenarios.
- The experimental results indicate that DiaHalu is a highly challenging task for most LLMs and existing detection methods, holding significant value for further research.

2 Related Work

2.1 Hallucination Detection Benchmarks

Hallucination predominantly centers on the inclination of large language models to generate nonsensical or untruthful contents for specific sources (Wang et al., 2023). Therefore, detecting hallucination in LLMs is of great importance. In recent years, researchers propose numerous benchmarks for the hallucination detection task.

In earlier years, hallucination detection benchmarks are primarily curated through manual methods or generated via conventional language models. FactCollect (Ribeiro et al., 2022) is an artificially generated, multi-source factual hallucination detection dataset. Muhlgay et al. (2023) collects error samples by instructing the language model based on pre-defined error types. HADES (Liu et al., 2022) constitutes a hallucination detection dataset produced by conventional language model BERT (Devlin et al., 2019). These benchmarks are not naturally generated by LLMs as in daily usage.

Consequently, some benchmarks are proposed to investigate the capacity for the direct generation of large language models. Zhao et al. (2023b); Fu et al. (2023); Chen et al. (2023a); Huang et al. (2023b); Zheng et al. (2023) enable LLMs to handle Question-Answer (QA) tasks and assess the factual accuracy of their responses. Concept-7 dataset used by Luo et al. (2023), evaluates whether a language model genuinely comprehends the meaning

[1] https://anonymous.4open.science/r/DiaHalu-A1E9/

of each concept, thereby determining the presence of hallucination. FactCHD (Chen et al., 2023c) is generated based on natural language text and knowledge graphs (KGs), deploying two major steps: Query and Response Data Simulation & Chain of Evidence and Human Filter, to detect fact-conflicting hallucination in LLMs. Mündler et al. (2023) employs a generative language model (gLM) to rewrite sentences according to the given context. New sentences compose a dataset that can evaluate whether the generated sentences exhibit knowledge-based self-contradiction hallucination. The aforementioned benchmarks mainly focus on detecting factuality hallucination (Guan et al., 2023), while ignoring the faithfulness hallucination (Huang et al., 2023a). The benchmark proposed in this paper extends to include faithfulness hallucination, that is, to evaluate the coherence and relevance of generated contents by LLMs.

Researchers also raise many sentence-level (Manakul et al., 2023; Zhao et al., 2023b; Guan et al., 2023) and passage-level (Yang et al., 2023; Feng et al., 2023b; Li et al., 2023) hallucination detection benchmarks. Nevertheless, the dialogue pattern holds broader applications within LLMs. While previous research does not, to our current knowledge, propose a dialogue-level hallucination detection benchmark. So, our DiaHalu is at the dialogue level.

2.2 Dialogue System in LLMs

It is the formidable dialogue capability exhibited by ChatGPT3.5 (Wu et al., 2023b) and GPT4 (OpenAI, 2023), that stimulates a heightened interest among researchers in large language models. The ability for multi-turn dialogue is among the most crucial features of LLMs, which has broader and more widespread applications in the current development of NLP.

Dialogue system can be primarily categorized into various types, including task-oriented dialogue (Wen et al., 2017), chit-chat dialogue (Sun et al., 2021), knowledge-grounded dialogue (Ghazvininejad et al., 2018), etc. In these scenarios, language models are required to exhibit long-term memory, recognize shifts in dialogue states, encompass extensive knowledge, and engage in reasoning processes (Huang and Chang, 2023). Despite the remarkable general capability of LLMs, there are still deficiencies in handling complex dialogue scenarios (Chen et al., 2017; Deriu et al., 2021).

Due to the critical significance of the dialogue ability, this paper underscores the detection of hallucination produced by LLMs in multi-turn dialogues.

3 The Overview of DiaHalu

3.1 Principles

The primary objective of this benchmark is to conduct hallucination detection in large language models. Hence, it is imperative to comprehend the meaning of *hallucination*. Hallucination predominantly centers on the inclination of LLMs to generate nonsensical or untruthful contents regarding specific sources (Wang et al., 2023). The significance of hallucination detection lies in elevating the quality of text generation, preventing misleading information and misunderstandings, supporting applications within professional domains, etc. Therefore, to enhance the universality of our benchmark, it encompasses various multi-turn dialogue scenarios and multiple types of hallucination, extending from factuality hallucination and faithfulness hallucination (Huang et al., 2023a).

3.2 Hallucination on Diverse Domains

We consider hallucination in diverse domains of multi-turn dialogue scenarios. Our benchmark covers a total of four domains, which are shown in Figure 2. Below are their specific descriptions.

Knowledge-grounded dialogue is designed for users to engage in knowledge-based dialogue with LLMs (Ghazvininejad et al., 2018). The knowledge includes world knowledge, factual knowledge, commonsense knowledge, and multi-hop web knowledge. It principally examines the accuracy of knowledge of the parameters in LLMs (Petroni et al., 2019).

Task-oriented dialogue is in a form of human-computer interaction, intending to accomplish a user-specified task (Wen et al., 2017). This type of dialogue system focuses on understanding the users' task requirements and utilizes a LLM to provide relevant information or perform specific tasks accordingly.

Chit-Chat dialogue involves open-ended and non-goal dialogues (Sun et al., 2021). We provide two LLMs with personas and facilitate a dialogue between them. This approach allows for the evaluation of their memory capabilities, conversational coherence, and relevance to the information being discussed.

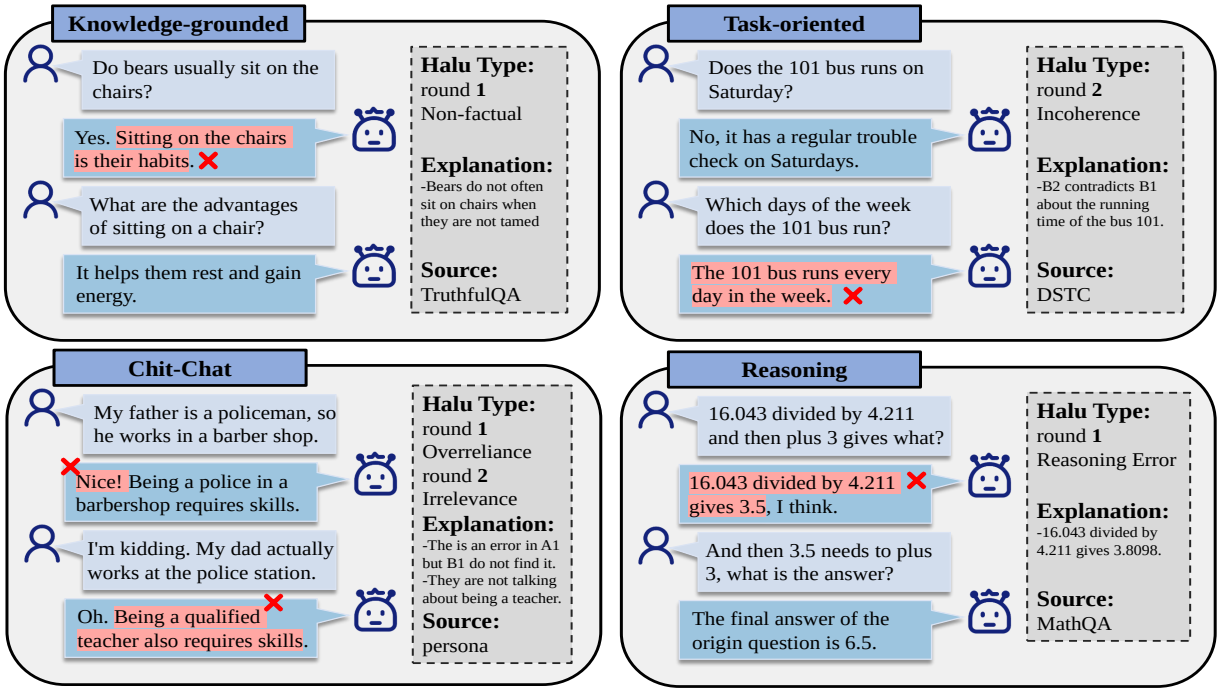

Figure 2: The demonstration of the DiaHalu benchmark, which covers four domains and five hallucination types within dialogue-level scenarios. We also provide explanations and sources in the benchmark.

Reasoning dialogue centralizes on the logical reasoning and understanding capabilities of LLMs. We have the models discuss mathematical problems to achieve the answers (Kakarla et al., 2024).

3.3 Hallucination Types

We consider factuality hallucination and faithfulness hallucination (Huang et al., 2023a). Based on Chen et al. (2023b), we carry out a detailed classification of faithfulness hallucination into Incoherence, Irrelevance, and Overreliance. Meanwhile, we introduce Reasoning Error within the reasoning dialogue. The example of each type can be referenced in Figure 2 and Appendix A.4.

Non-factual implies whether it aligns with factual information.

Incoherence includes input-conflicting, context-conflicting and self-conflicting contents in the dialogue.

Irrelevance means that something unrelated to the topic of the conversation comes up.

Overreliance is that the LLM excessively relies on the correctness of context, generating responses for unanswerable content (Slobodkin et al., 2023).

Reasoning Error covers all errors within the reasoning dialogue.

4 The Construction of DiaHalu

4.1 The Collection of Dialogue Topics

Since we confirm four domains for DiaHalu, the first step is to collect the topics for each dialogic domain.

For **knowledge-grounded dialogue**, we take into account world knowledge, factual knowledge, commonsense knowledge and multi-hop web knowledge. Therefore, we gather dialogue topics for TruthfulQA (Lin et al., 2022), CommonsenseQA (Talmor et al., 2019) and CWQ (Talmor and Berant, 2018) datasets. There are also topics provided by GPT4 (OpenAI, 2023) and social media (including the authors). As for **task-oriented dialogue**, we primarily apply the most widely used MultiWOZ (MultiWOZ 2.1) (Budzianowski et al., 2018) which has 7 real-life scenarios. To enrich the dialogue settings, we also consider the DSTC (DSTC 1.0) (Williams et al., 2013) dataset with a focus on bus routes. GPT4 and social media are harnessed to augment user behaviors and generate more dialogue occasions. We define the LLMs with personas primarily from Jandaghi et al. (2023) and facilitate an open **chit-chat dialogue** between

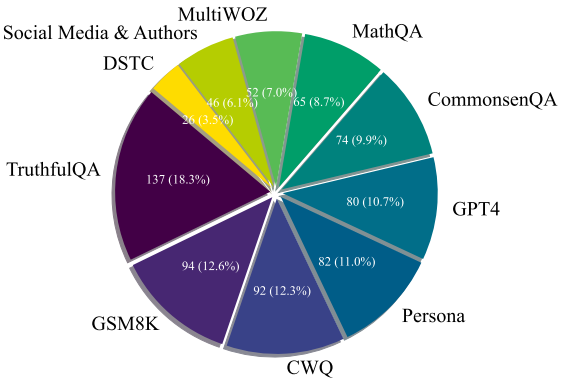

Figure 3: The distribution of the 10 sources for dialogue topics.

Attribute	Attribute Value
Benchmark Name	DiaHalu
Generated by	ChatGPT3.5
Sample Form	dialogue-level
Sample Numbers	748
Dialogue Rounds	6-10
Avg. Rounds	6.8951
Domain Numbers	4
Hallucination Types	5
Max. Response Length (Words)	183
Avg. Response Length (Words)	13.5091

Table 1: The statistical information of the benchmark.

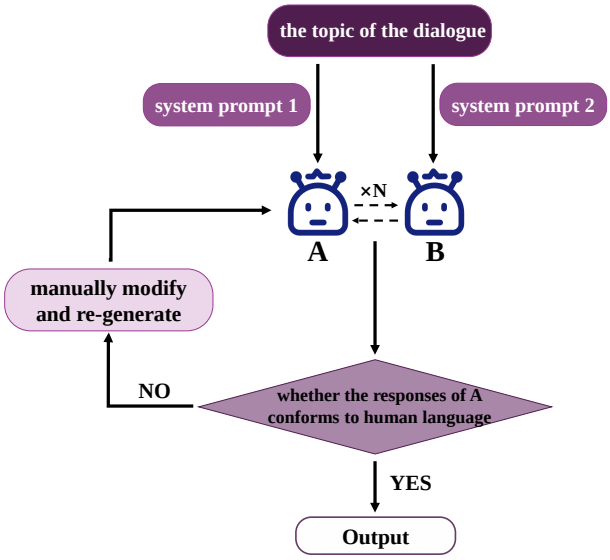

Figure 4: The complete process of dialogue generation.

them. Additionally, we make use of mathematical problems to assess the logical **reasoning dialogue** ability of LLMs. These problems are sourced from GSM8K (Cobbe et al., 2021) and MathQA (Amini et al., 2019), both of which involve mathematical problems and solving processes encountered by middle school students. The overall distribution of the above sources for dialogue topics is illustrated in Figure 3.

4.2 Dialogue Generation

Once finishing collecting the dialogue topics for each domain, we leverage ChatGPT3.5 to generate conversations in the format of self-dialogue. The complete process of dialogue generation is illustrated in Figure 4.

Initially, we integrate the dialogue topics into two system prompts, which are then inputted separately into two ChatGPT3.5. These two system prompts guide the LLMs to generate N rounds of dialogue in a given domain and topic. More details of the system prompts can be found in Appendix A.1. Then, for knowledge-grounded dialogue and task-oriented dialogue, we manually examine all responses from A to determine their adherence to human language. For in both scenarios, we consider real human-machine interaction, aiming to assess the LLMs' adaptability to genuine human behaviors. (We assume that A is the user and B is the LLM. In this setup, we ensure the accuracy of A and only annotate the contents of B.) The chit-chat dialogue and reasoning dialogue are relatively unconstrained and freely conducted, without incorporating any specific human-machine interaction settings (Section 3.2). They necessitate only their memory and comprehension capabilities regarding contextual information, thereby minimizing the need for manual intervention. Consequently, for the responses of A in knowledge-grounded and task-oriented scenarios, where the contents do not conform to human language, we manually modify and have LLMs re-generate. Eventually, we obtain the complete dataset of multi-turn dialogues.

We provide one output sample in Appendix A.2. The comparison with other benchmarks is in Appendix A.3, highlighting the distinctive features that our DiaHalu is a naturally generated dialogue-level benchmark by LLMs, with various kinds of hallucination and explanations. The statistical information of the whole benchmark is in Table 1.

4.3 Human Annotation

Annotating the hallucination and its types in the benchmark is a very challenging task. Since there may be more than one instance of hallucination

in multi-turn dialogues. Also, some hallucination types in edge cases are difficult to differentiate. Therefore, the entire annotation process demands a high level of expertise from annotators and requires detailed definitions for ambiguous contents.

The annotators of our benchmark are all seasoned researchers in the field of linguistics and natural language processing. The annotators all obtain at least a bachelor's degree, pass the IELTS or TOEFL exams, and are proficient in using search engines such as Google and Bing. In addition, everyone exhibits strong collaborative and communicative skills. To ensure the validity of the annotation information, we invite experienced experts in the field of LLMs' hallucination detection from both academia and industry to conduct random sampling checks.

Annotation process is divided into three steps. (1) Each annotator labels 50 samples for each domain, followed by a careful discussion. The annotators are required to label the presence of hallucination, hallucination types and locations, along with the corresponding explanations. The discussions intricately define the application scope of each hallucination label (Appendix A.4), and ultimately we decide to annotate the corresponding hallucination type on the first dialogue round in which hallucination occurs. (2) All the annotators label the entire dataset, discussions and corrections are made for inconsistent annotations. (3) Statistical analysis is performed on the annotated results. For more details, please refer to Appendix A.5.

Annotation Consistency For evaluating the inter-annotator consistency, we calculate the Fleiss's Kappa (Randolph, 2005) of Inter-Annotator Agreement (IAA) (Artstein, 2017), which is a statistical measure used to assess the degree of agreement among multiple raters for a set of items.

After the whole annotation process, we achieve a label matrix $\mathbf{L} \in \mathbb{R}^{N_s * N_A}$. N_s, N_A represent the number of dialogue samples and annotators respectively. The calculation for Fleiss's Kappa is shown below:

$$P_e = \left(\frac{\sum_{i=1}^{N_s} \sum_{j=1}^{N_A} \chi_{\{0\}}(\mathbf{L}[i,j])}{N_s * N_A} \right)^2 + \left(\frac{\sum_{i=1}^{N_s} \sum_{j=1}^{N_A} \chi_{\{1\}}(\mathbf{L}[i,j])}{N_s * N_A} \right)^2 \quad (1)$$

$$P_o = \frac{1}{N_s} \sum_{i=1}^{N_s} \frac{\left(\sum_j^{N_A} \chi_{\{0\}}(\mathbf{L}[i,j])\right)^2}{N_A * (N_A - 1)} + \frac{\left(\sum_j^{N_A} \chi_{\{1\}}(\mathbf{L}[i,j])\right)^2 - N_A}{N_A * (N_A - 1)} \quad (2)$$

$$Fleiss's Kappa = \frac{P_o - P_e}{1 - P_e} \quad (3)$$

where P_o and P_e represent the relative observed agreement among annotators and the hypothetical probability of chance agreement respectively. $\chi_y(Y)$ is the Indicator Function, which means when the value of Y is in set y, the whole function equals 1.

According to the above formulas, the calculated result for Fleiss's Kappa of our benchmark is 0.8709, representing almost perfect agreement among all the annotators.

Annotation Results After annotating the entire dataset, we conduct several statistical analyses on it. Table 2 reveals the probability of hallucination occurring in each dialogue domain. The results indicate that hallucinations are highly likely to arise in Knowledge-grounded dialogue and reasoning dialogue. Therefore, the knowledge and reasoning abilities of LLMs still need further improvement. Despite LLMs' powerful multi-turn dialogue capability, faithfulness hallucination such as irrelevance, incoherence and overreliance still persists. Figure 5 presents the proportion of each hallucination type in each dialogue domain. Irrelevance, incoherence, and overreliance Widely exit in daily dialogue contexts, such as task-oriented and chit-chat scenarios. In knowledge-grounded dialogue, the factuality hallucination constitutes a significant proportion, while in reasoning dialogue, almost all hallucinations are defined as errors in reasoning. This statistical information can help us understand the types of hallucination in LLMs's multi-turn dialogues, facilitating an exploration of their origins and contributing to the elimination of these types of hallucination.

5 Experiments

In this section, we assess the performance of several evaluation models and methods on the dataset we introduced. Thereby, we can determine the effectiveness of existing methods in detecting dialogue-level hallucination.

	Knowledge	Task	Chit	Reasoning	Overall
# Number	272	131	186	159	748
# Non-Halu	141	84	115	79	419
# Halu	131	47	71	80	329
Halu Rate (%)	48.16	35.88	38.17	50.31	43.98

Table 2: The statistical information of hallucination on the four dialogue domains.

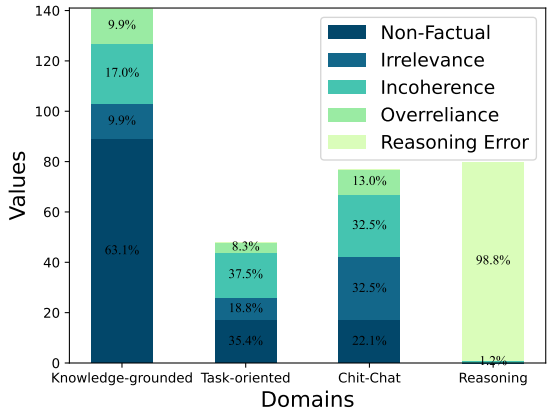

Figure 5: The distribution of five different hallucination types within the four dialogue domains.

5.1 Baselines

We select some powerful LLMs by providing specific prompts, to assess the presence of hallucinations. Similarly, we also conduct experiments using existing hallucination detection methods.

Random A straightforward approach that randomly generates a label for each sample.

SelfCheckGPT (Manakul et al., 2023) It is a widely used black-box hallucination detection framework. It rephrases the contents to be detected while ensuring the consistency of semantics by LLMs with different temperatures. Furthermore, it calculates the consistency between the original and the rephrased contents using five methods, thereby determining whether there exits hallucination. The indices B, N and P respectively denote scoring with BERTScore (Zhang et al., 2020), scoring with Natural Language Inference methods (He et al., 2023) and the direct judgment using prompts.

LLaMA-30B && Vicuna-33B They are two well-pretrained and widely deployed open-source LLM backbones[2] (Touvron et al., 2023; Wei-Lin Chiang, 2023). We provide a specially designed prompt to assist with detection. More details are shown in Appendix A.6.

ChatGPT3.5 && ChatGPT4 Both of these models are developed by OpenAI[3]. ChatGPT3.5 marks the beginning of the era of large language models and GPT4 is currently the most powerful language model (Wu et al., 2023b; OpenAI, 2023). We employ the same prompt as LLaMA and Vicuna do.

5.2 Metrics

Despite we annotating the types of hallucinations in the dataset, achieving consistent labels even among humans requires further discussion. Therefore, similar to past hallucination detection efforts, we merely focus on a binary classification task of determining the existence of hallucination. Consequently, we utilize binary classification evaluation metrics. The positive label for this classification task is set as "Halu", for our main focus is testing the model's ability to recognize hallucination.

5.3 Main Results

From the results in Table 3, we can get the following conclusions.

First, DiaHalu is a very challenging benchmark for dialogue-level hallucination detection. Except for ChatGPT4, the F1 scores of all other detection methods do not exceed 50.00. Existing language models, such as LLaMA-30B and Vicuna-33B, are not effective in accurately discerning most samples that involve hallucinations. Regarding the specialized detection method SelfCheckGPT, it achieves relatively better performance when applying the prompt and NLI methods. However, distinguishing between positive and negative samples proves challenging with BERTScore.

Second, when performing self-detection using ChatGPT3.5, there is a noticeable phenomenon of overconfidence. Our dataset is primarily generated by ChatGPT3.5, which exhibits high confidence in its output. Despite providing a specially designed detection prompt, it still struggles to differentiate whether the dialogue content is hallucinating or not. So, the majority of its output labels are "Non-Halu".

Third, the faithfulness hallucination is more difficult to detect for LLMs. Apart from the spe-

[2]https://huggingface.co/huggyllama/llama-30b, https://huggingface.co/lmsys/vicuna-33b-v1.3

[3]https://chat.openai.com/

Method	Knowledge-grounded			Task-oriented			Chit-Chat			Reasoning			Overall		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Random	48.19	30.53	37.38	35.71	21.28	26.67	32.50	18.31	23.42	50.00	23.75	32.20	43.39	24.92	31.66
SelfCheckGPT$_B$	47.92	17.56	25.70	32.14	19.15	24.00	37.93	15.49	22.00	57.50	28.75	38.33	45.52	20.06	27.85
SelfCheckGPT$_N$	64.41	29.01	40.00	34.25	53.19	41.67	46.58	47.89	47.22	64.00	20.00	30.48	49.13	34.35	40.43
SelfCheckGPT$_P$	56.45	26.72	36.27	44.73	36.17	40.00	43.33	54.92	**48.45**	60.96	62.50	61.73	51,84	42.86	46.92
LLaMA-30B	38.46	3.82	6.94	40.00	4.26	7.69	50.00	11.27	18.39	71.43	6.25	11.49	48.78	6.08	10.81
Vicuna-33B	50.00	4.58	8.39	100.00	4.26	8.16	28.57	2.82	5.13	41.38	15.00	22.02	44.00	6.69	11.61
ChatGPT3.5	50.00	0.76	1.50	0.00	0.00	0.00	42.86	4.23	7.69	55.56	6.25	11.24	45.00	2.74	5.16
ChatGPT4	82.46	35.88	**50.00**	73.68	29.79	**42.42**	56.52	18.31	27.66	71.43	81.25	**76.02**	73.16	42.25	**53.56**

Table 3: The classification results on four kinds of baselines, and the best F1 scores are in bold form.

Method	Knowledge-grounded			Task-oriented			Chit-Chat			Reasoning			Overall		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
ChatGPT3.5	50.00	0.76	1.50	0.00	0.00	0.00	42.86	4.23	7.69	55.56	6.25	11.24	45.00	2.74	5.16
w/ CoT	45.45	3.82	7.04	20.00	2.13	3.85	28.57	2.82	5.13	53.85	8.75	15.05	41.67	4.56	8.22
w/ retrieval	70.00	5.34	9.93	-	-	-	-	-	-	71.43	12.50	21.27	-	-	-
ChatGPT4	82.46	35.88	50.00	73.68	29.79	42.42	56.52	18.31	27.66	71.43	81.25	76.02	73.16	42.25	53.56
w/ CoT	82.02	55.73	66.36	72.41	44.68	55.26	69.05	40.85	51.33	69.52	91.25	78.92	73.96	59.57	65.99
w/ retrieval	75.90	48.09	58.88	-	-	-	-	-	-	70.65	81.25	75.58	-	-	-

Table 4: The results of CoT and retrieval techniques which can enhance hallucination detection on ChatGPT3.5 and ChatGPT4.

cialized hallucination detection method SelfCheckGPT, the results from directly harnessing LLMs for judgment indicate that the recognition accuracy for task-oriented and chit-chat domains of dialogue are much lower than that for the knowledge-grounded and reasoning dialogue. This is because the hallucination types in the knowledge-grounded and reasoning dialogue are primarily Non-factual and Reasoning Error, which present in a more direct and apparent manner. While task-oriented and chit-chat domains mainly consist of three types of faithfulness hallucination, which requires a LLM to possess long-term memory and the ability to recognize topic transitions in dialogues.

5.4 Chain-of-Thought and Retrieval for Detection

Chain-of-Thought and Retrieval are two important techniques for enhancing the reasoning ability of LLMs. Chain-of-Thought (CoT) (Feng et al., 2023a) describes the organized sequence of logical reasoning that unfolds during thinking. Retrieval (Gao et al., 2023) means retrieving relevant contents from the media to supplement external knowledge for LLMs. In this section, we test whether these two techniques can improve the effectiveness of hallucination detection in Table 4. More details are depicted in Appendix A.7.

The experimental results indicate that both methods have facilitating effects on hallucination detection. However, the combination of GPT4 with retrieval methods shows a decrease of around 0.5 points in F1 score in the domain of reasoning detection. We believe that GPT4 inherently possesses a strong reasoning ability, and therefore, additional retrieval contents introduce noise to its judgments.

6 Conclusion

In this paper, we propose a dialogue-level hallucination evaluation benchmark. We construct the benchmark in a three-step process and have the experts annotate all the data. DiaHalu covers four multi-turn dialogue domains and five hallucination subtypes. Experiments through some well-known LLMs and detection methods on the dataset show that DiaHalu is a challenging task, holding significant value for further research (Appendix A.8).

Limitations

This paper proposes a novel dialogue-level hallucination detection evaluation benchmark named DiaHalu. The benchmark covers four multi-turn dialogue domains and five hallucination subtypes. There is significant value for further research. However, two main limitations also exist. (1) During the second step of the benchmark construction phase, aligning the contents of speaker A with human language consumes a considerable amount of time and effort. Frequent calls to the ChatGPT3.5 API_Keys result in a significant expense. Simultaneously, achieving consistency among all annotators led to prolonged discussion time. (2) We do not partition the dataset into training, validation, and test sets. This is because we consider the black-box detection scenario for those closed-source LLMs. If a division into these three types of datasets is required, it would necessitate more data samples and larger resource consumption.

Ethics Statement

The benchmark is primarily generated by ChatGPT3.5. We obtain all the API_Keys through a paid subscription. All the annotators are real and they receive corresponding compensation and rewards. The entire process and outcomes are free from intellectual property and ethical legal disputes.

Acknowledgements

Due to the constrictions of the double-blind review mechanism, we will acknowledge all contributors and institutions as soon as our paper is accepted.

References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms.

Yihao Ang, Qiang Huang, Anthony K. H. Tung, and Zhiyong Huang. 2023. A stitch in time saves nine: Enabling early anomaly detection with correlation analysis. In *39th IEEE International Conference on Data Engineering, ICDE 2023, Anaheim, CA, USA, April 3-7, 2023*, pages 1832–145. IEEE.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

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

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor.*, 19(2):25–35.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023a. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6325–6341. Association for Computational Linguistics.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023b. FELM: benchmarking factuality evaluation of large language models. *CoRR*, abs/2310.00741.

Xiang Chen, Duanzheng Song, Honghao Gui, Chengxi Wang, Ningyu Zhang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2023c. Unveiling the siren's song: Towards reliable fact-conflicting hallucination detection. *CoRR*, abs/2310.12086.

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

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.*, 54(1):755–810.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). *CoRR*, abs/2307.02046.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023a. Towards revealing the mystery behind chain of thought: a theoretical perspective. *CoRR*, abs/2305.15408.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023b. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 933–952. Association for Computational Linguistics.

Xue-Yong Fu, Md. Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023. Are large language models reliable judges? A study on the factuality evaluation capabilities of llms. *CoRR*, abs/2311.00681.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5110–5117. AAAI Press.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *CoRR*, abs/2310.14564.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1049–1065. Association for Computational Linguistics.

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

Yuheng Huang, Jiayang Song, Zhijie Wang, Huaming Chen, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *CoRR*, abs/2307.10236.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *CoRR*, abs/2312.10007.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Sanjit Kakarla, Danielle Thomas, Jionghao Lin, Shivang Gupta, and Kenneth R. Koedinger. 2024. Using large language models to assess tutors' performance in reacting to students making math errors. *CoRR*, abs/2401.03238.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. 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 6723–6737. Association for Computational Linguistics.

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

Junyu Luo, Cao Xiao, and Fenglong Ma. 2023. Zero-resource hallucination prevention for large language models. *CoRR*, abs/2309.02654.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*,

pages 9004–9017. Association for Computational Linguistics.

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. *CoRR*, abs/2307.06908.

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

Jesutofunmi A. Omiye, Haiwen Gui, Shawheen J. Rezaei, James Zou, and Roxana Daneshjou. 2023. Large language models in medicine: the potentials and pitfalls. *CoRR*, abs/2309.00087.

OpenAI. 2023. Gpt-4 technical report.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. Factgraph: Evaluating factuality in summarization with semantic graph representations. 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 3238–3253. Association for Computational Linguistics.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory unanswerablity: Finding truths in the hidden states of over-confident large language models. *CoRR*, abs/2310.11877.

Elior Sulem, Jamaal Hay, and Dan Roth. 2021. Do we know what we don't know? studying unanswerable questions beyond squad 2.0. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4543–4548. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1570–1583. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 641–651. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Jiayang Cheng, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Zi Lin Ying Sheng Zhanghao Wu Hao Zhang Lianmin Zheng Siyuan Zhuang Yonghao Zhuang Joseph E. Wei-Lin Chiang, Zhuohan Li. 2023.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 438–449. Association for Computational Linguistics.

Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*, pages 404–413. The Association for Computer Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023a. Bloomberggpt: A large language model for finance. *CoRR*, abs/2303.17564.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023b. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

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, Singapore, December 6-10, 2023*, pages 3898–3908. Association for Computational Linguistics.

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

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, 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. 2023a. A survey of large language models. *CoRR*, abs/2303.18223.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2023b. Knowing what llms DO NOT know: A simple yet effective self-detection method. *CoRR*, abs/2310.17918.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

A Appendices

A.1 System Prompts for Dialogue Generation

In this section, we present the specific form of the system prompts for the four dialogue domains. We use ChatGPT3.5 (gpt-3.5-turbo-1106)[4] with the temperature 0.1. The brief system prompts of the knowledge-grounded dialogue, task-oriented dialogue, chit-chat dialogue and reasoning dialogue are respectively presented in Figure 6, Figure 7, Figure 8 and Figure 9.

A.2 Data Format

We provide the specific format of one sample from the benchmark in Figure 10.

A.3 The Comparison with Other Benchmarks

In Table 5, we present the differences between DiaHalu and other hallucination detection benchmarks, highlighting the distinctive features that our DiaHalu is a naturally generated dialogue-level benchmark by LLMs, with various kinds of hallucination and explanations. All the compared benchmarks can be referred to in Section 2.1. HaluEval, WikiBio+, and PHD benchmark come from the paper Li et al. (2023), Manakul et al. (2023) and Yang et al. (2023) respectively.

A.4 The Application Scope of Hallucination Labels

Annotating the hallucination and its types in the benchmark is a very challenging task. One of the reasons is that some hallucination types in side cases are defiant to differentiate. Therefore, in the first stage of annotation, we provide detailed definitions for each hallucination type.

Non-factual implies that it does not align with facts or introduce elements that do not exist in real life.

Incoherence means there is a contradiction between the dialogue role and his context, a contradiction with the dialogue object, or an internal contradiction within the current response (in a declarative sentence). It also refers to factual and relevant nonsense or contextually inappropriate responses.

Irrelevance involves responses that are irrelevant to the dialogue topic or are unrelated due to a misunderstanding (in the case of a question).

Overreliance is that the LLM excessively relies on the correctness of context, generating responses for unanswerable content (in a declarative sentence).

[4] https://platform.openai.com/docs/guides/text-generation/chat-completions-api

system prompt 1:
Your task is to generate a response based on the above conversation contents and the topic [TOPIC]. Your tone and way of thinking should be as similar to humans as possible. It is a part of a multi-round knowledge-grounded dialogue between A and B.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response could be diverse and colorful, covering various fields such as literature, science and engineering, etc.
3. The response to be semantically rich and grammatically varied.
4. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
5. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
6. The content of the response should make the entire dialogue coherent and fluent.
7. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.

---

system prompt 2:
###Your task is to generate a response based on the above conversation contents and the topic [TOPIC]. It is a part of a multi-round knowledge-grounded dialogue between A and B.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response could be diverse and colorful, covering various fields such as literature, science and engineering, etc.
3. The response to be semantically rich and grammatically varied.
4. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
5. Each response must be strongly related to the previous responses, the generation of dialogue can reference the history of the conversation.
6. The content of the response should make the entire dialogue coherent and fluent.
7. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.

Figure 6: The brief system prompts for knowledge-grounded dialogue.

Benchmark	By LLMs	Faith Halu	Dialogue	Explanation
FactCollect	-	-	-	-
HADES	-	✓	-	✓
FactCHD	✓	-	-	✓
HaluEval	-	-	✓	✓
WikiBio+	✓	-	-	-
PHD	✓	-	-	-
Ours	✓	✓	✓	✓

Table 5: The comparison between our DiaHalu and other benchmarks. 'By LLMs', 'Faith Halu', 'Dialogue', and 'Explanation' mean whether it is naturally generated by LLMs, whether it provides faithfulness hallucination, whether it is at dialogue level, and whether there are explanations respectively.

Reasoning Error covers all errors within the reasoning dialogue.

A.5 The Supplementary Details for Annotation

We use Label Studio [5] for labeling, which is an online open-source data labeling platform in the field of artificial intelligence. The annotation interface is depicted in Figure 11.

[5]https://labelstud.io/

A.6 The Prompt Designed for Detection

In Figure 12, we show the whole prompt specially designed for hallucination detection of the baselines: LLaMA-30B, Vicuna-33B, ChatGPT3.5, and ChatGPT4. It is worth noting that due to the poor instruction-following ability and the disorderly output format of the LLaMA and Vicuna models, we conduct experiments in a 1-shot manner. The ChatGPT3.5 used for detection is ChatGPT3.5 (gpt-3.5-turbo-1106)[6] and the ChatGPT4 is ChatGPT4 (gpt-4-1106-preview) [7].

A.7 The Settings for CoT and Retrieval

We employ CoT in all four domains to enhance the performance of ChatGPT3.5 and ChatGPT4. The specific CoT is illustrated in Figure 13. While only knowledge-grounded and reasoning domains are tested with retrieval via Google[8]. This is because domains of task-oriented and chit-chat mainly involve scenarios related to daily life or virtual worlds, without specific domain knowledge as supplementary information.

A.8 Future Works

Dialogue-level hallucination detection is an important work in the future. We propose the first dialogue-level hallucination detection evaluation

[6]https://platform.openai.com/docs/guides/text-generation/chat-completions-api

[7]https://openai.com/gpt-4

[8]https://console.cloud.google.com/apis/library

system prompt 1:
Your task is to generate a response based on the above conversation contents and the topic [TOPIC]. Your tone and way of thinking should be as similar to customers as possible. It is a part of a multi-round task-oriented dialogue between A and B. And here are some related action of the you (which can be mentioned in the dialogue): [ACTIONS].

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response could be diverse and colorful, covering various fields such as literature, science and engineering, etc.
3. The response to be semantically rich and grammatically varied.
4. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
5. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
6. The content of the response can be creative or personalized, design vague questions, quick topic switching, and repeated questions.
7. Give specific number information.
8. Use your imagination and global knowledge should also be included in the conversation.
9. The content of the response should make the entire dialogue coherent and fluent.
10. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.

---

system prompt 2:
###Your task is to generate a response based on the above conversation contents and the topic [TOPIC]. Your tone and way of thinking should be as similar to servers as possible. It is a part of a multi-round task-oriented dialogue between A and B.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response could be diverse and colorful, covering various fields such as literature, science and engineering, etc.
3. The response to be semantically rich and grammatically varied.
4. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
5. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
6. the service provider can refuse the customer's request, relying on your imagination to give the reason. The service provider can ask the customer for specific information.
7. Give specific number information.
8. Use your imagination and global knowledge should also be included in the conversation.
9. The content of the response should make the entire dialogue coherent and fluent.
10. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.

Figure 7: The brief system prompts for task-oriented dialogue.

benchmark and experimental results show that it is a very challenging task. Therefore, developing methods based on this dataset to achieve a relatively high recognition accuracy is highly valuable.

Dialogue-level hallucination elimination is an extension task of this work. Most existing hallucination elimination methods primarily focus on passage-level factuality hallucination (Luo et al., 2024). Hallucination elimination at the dialogue level not only requires the models to have a long-context memory capability but also demands the abilities to recognize changes in topics and logical transitions in the dialogue. These abilities are also helpful in addressing faithfulness hallucinations. At the same time, improving the accuracy of knowledge in the LLM parameters and the reasoning ability are equally important.

Hallucination snowballing is the phenomenon that LLMs tend to accumulate hallucinations rather than self-correcting during the generation process (Zhang et al., 2023). Some previous works also validate this phenomenon (Azaria and Mitchell, 2023; Ang et al., 2023). In our dataset, there is a noticeable issue of hallucination snowballing, so we only label the first occurrence of hallucination in our dataset. Eliminating hallucination snowballing is also important in the future.

The Unanswerability of LLMs During the annotation process of this dataset, we introduce a hallucination category termed "overreliance", which represents answering unanswerable content (Slobodkin et al., 2023; Sulem et al., 2021). This phenomenon signifies that LLMs tend to trust the input provided by users. Sometimes, even when there are errors in user input, the LLMs still fail to recognize them. A few past researches explore the related areas. However, the issue in the application of human-machine interaction and multi-agent scenarios remains crucial.

system prompt 1:
Your task is to generate a response based on the above conversation contents and your personas [PERSONA1]. Your tone and way of thinking should be as similar to personas as possible. It is a part of a multi-round chit-chat dialogue between A and B.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response could be diverse and colorful, covering various fields such as literature, science and engineering, etc.
3. The response to be semantically rich and grammatically varied.
4. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
5. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
6. The content of the response can be creative or personalized, design vague questions, quick topic switching, and repeated questions.
7. Give specific number information.
8. Use your imagination and global knowledge should also be included in the conversation.
9. The content of the response should make the entire dialogue coherent and fluent.
10. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.

---

system prompt 2:
Your task is to generate a response based on the above conversation contents and your personas [PERSONA2]. Your tone and way of thinking should be as similar to personas as possible. It is a part of a multi-round chit-chat dialogue between A and B.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response could be diverse and colorful, covering various fields such as literature, science and engineering, etc.
3. The response to be semantically rich and grammatically varied.
4. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
5. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
6. The content of the response can be creative or personalized, design vague questions, quick topic switching, and repeated questions.
7. Give specific number information.
8. Use your imagination and global knowledge should also be included in the conversation.
9. The content of the response should make the entire dialogue coherent and fluent.
10. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.

Figure 8: The brief system prompts for chit-chat dialogue.

system prompt 1:
Your task is to generate a response based on the above conversation contents and the problem [PROBLEM]. It is a part of a multi-round reasoning dialogue between A and B. You need to discuss your thoughts step by step to get the answer.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
3. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
4. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.
5. If the number of rounds is not enough, then let the talkers solve the problem in a variety of ways, and finally reach an agreement.
6. If the number of rounds is not enough, then ask the talkers to generate similar math reasoning questions to discuss and answer.
7. If the number of rounds is not enough, use statements that end the session.
8. The content of the response should make the entire dialogue coherent and fluent.

---

system prompt 2:
Your task is to generate a response based on the above conversation contents and the problem [PROBLEM]. It is a part of a multi-round reasoning dialogue between A and B. You need to discuss your thoughts step by step to get the answer.

###The response must adhere to the following requirements:
1. The response must be within 3 sentences.
2. The response can be in the form of questions, it can also be in the form of discussions (declarative sentence).
3. Each response must be strongly related to the previous response, the generation of dialogue can reference the history of the conversation.
4. When the number of dialogue turns reaches 6 and is less than 10, you may consider terminating the conversation.
5. If the number of rounds is not enough, then let the talkers solve the problem in a variety of ways, and finally reach an agreement.
6. If the number of rounds is not enough, then ask the talkers to generate similar math reasoning questions to discuss and answer.
7. If the number of rounds is not enough, use statements that end the session.
8. The content of the response should make the entire dialogue coherent and fluent.

Figure 9: The brief system prompts for reasoning dialogue.

{
ID: 1,
Dialogue: “A1: Tell me about the solar system. \n B1: The solar system consists of the sun, planets, moons, asteroids, and comets orbiting around the sun. \n A2: Tell me about Pluto. \n B2: Pluto is a dwarf planet in the Kuiper Belt, and it used to be considered the ninth planet in our solar system before being reclassified in 2006. \n A3: Which is the eighth planet? \n B3: Neptune. \n A4: What is the distance between Neptune and Pluto? \n B4: The distance between Neptune and Pluto can vary as both planets have elliptical orbits, but on average, it's about 4.5 billion kilometers (2.8 billion miles). \n A5: What is the distance between Sun and Pluto? \n B5: The distance between the Sun and Pluto also varies due to their elliptical orbits. On average, it's about 5.9 billion kilometers (3.67 billion miles). \n A6: Which is closer from Neptune? Sun or Pluto? \n B6: The Sun is closer to Neptune than Pluto. \n A7: Whether Pluto is in the solar system? \n B7: Yes, Pluto is in the solar system. ”,
Label: 1,
Hallucination Type: [[7,"Incoherence"]],
Domain: “Knowledge-grounded”,
Explanation: ["B7 contradicts B2"],
Source: “Social Media”
}

Figure 10: One sample from the benchmark

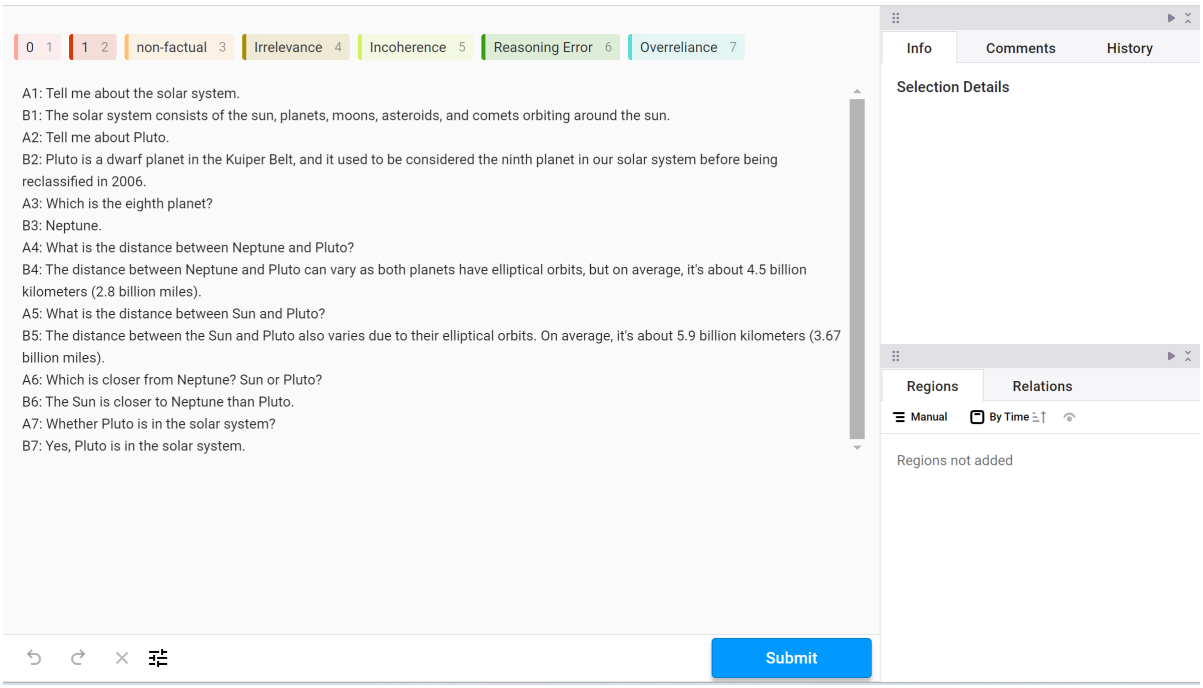

Figure 11: The annotation interface on Label Studio.

#Role:
Please tell me if there are any errors in the multi-turn dialogue I gave you.

#Some explanations about the multi-turn dialogue I input:
1. This is a multi-turn dialogue that occurs between A and B.
2. The multi-turn dialogue is [TYPE] type dialogue.
3. If the conversation type is either casual or reasoning-oriented, A and B are two people. In other cases, A is a person and B is ChatGPT, i.e., artificial intelligence.
4. Now, our subject for analysis is X, who can be either A, B, or both. X's conversational partner is another person. I want you to check for errors in [OBJECTS]。
5. Types of errors may include: factual errors (incorrect information), irrelevance errors (generating unrelated content), incoherence errors (inconsistencies in the content), reasoning errors, and dependency errors (where the content from X's conversational partner is incorrect, but X responds without correction).
6. I will input each multi-turn dialogue gradually. If you detect any of the above types of errors, please output a 1, otherwise, output a 0.
7. Please judge each sentence carefully and check each sentence against its historical context.
8. Note that there is at least a 40% chance that the dialogues I provide will contain the aforementioned errors. Please help me carefully and thoroughly check them.

#Skills:
Please read the multi-turn dialogue I provide carefully.
Then output either 0 or 1, where 0 means no errors, and 1 means there are errors. Please do not output any other unrelated content. Just output either 0 or 1.
However, if I ask you to provide an explanation, please do so.

Figure 12: The whole prompt for hallucination detection of the baselines.

Hallucination types may include:
1. Factual (errors in facts, mentioning things that do not exist in the real world)
2. Irrelevance (producing unrelated content)
3. Incoherence (inconsistency errors, contradictions between the generated content, including contradictions within the content generated for X, contradictions in the dialogue history of X, or contradictions between the content of X and the content of the entity X is conversing with)
4. Reasoning errors
5. Dependency errors (where the content of the entity X is incorrect, but X still responds without correction).

So please:
1. Check whether each response aligns with world knowledge, factual knowledge or common knowledge.
2. Verify whether each response is relevant to the overall conversation topic or the current context of the dialogue.
3. Examine whether the dialogue content contradicts with the preceding or following sentences or within itself.
4. Check for logical errors.
5. Examine whether each dialogue responds to incorrect content.

Figure 13: The whole CoT for the four domains of dialogue.