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

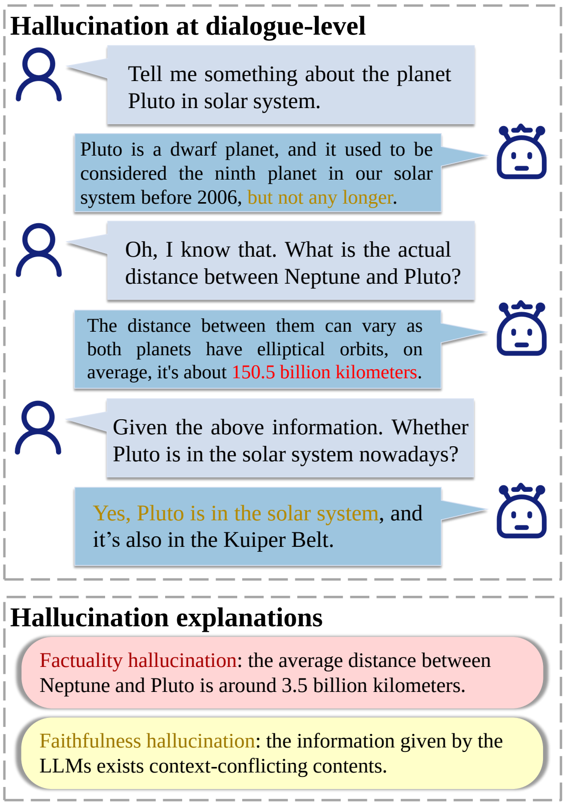
Anonymous ACL submission

Abstract

Since large language models (LLMs) achieve significant success in recent years, the hallucination issue remains a challenge, and numerous benchmarks are proposed for hallucination detection. 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 dedicated dialogue-level hallucination evaluation benchmark for LLMs to our knowledge. Initially, we integrate the collected topics into system prompts and facilitate a dialogue between two LLMs. 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., 2023b), addressing many downstream tasks through natural language generation (NLG) technique, achieve significant success across diverse fields of natural language processing (NLP) (Li et al., 2024a; 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., 2023b).



Tell me something about the planet Pluto in solar system.

Pluto is a dwarf planet, and it used to be considered the ninth planet in our solar system before 2006, but not any longer.

Oh, I know that. What is the actual distance between Neptune and Pluto?

The distance between them can vary as both planets have elliptical orbits, on average, it's about 150.5 billion kilometers.

Given the above information. Whether Pluto is in the solar system nowadays?

Yes, Pluto is in the solar system, and it's also in the Kuiper Belt.

Factuality hallucination: the average distance between Neptune and Pluto is around 3.5 billion kilometers.

Faithfulness hallucination: the information given by the LLMs exists context-conflicting contents.

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

Despite many advantages of large language models, the issue of hallucination remains a primary concern (Ji et al., 2023; Zhang et al., 2023c). Hallucination predominantly centers on the inclination of LLMs to generate nonsensical or untruthful contents for specific sources (Wang et al., 2023a). Therefore, the occurrence of hallucination poses risks to the application of large language models in various real-world scenarios (Omiye et al., 2023; Wu et al., 2023b; Chen et al., 2024).

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., 2023b; 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., 2023b), 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 propose many sentence-level (Manakul et al., 2023; Zhao et al., 2023c) 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 rarely mentioned in previous researches, 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 dialogue from various sources, incorporate the topics into artificially designed system prompts and input them into two LLMs, enabling them to engage in a multi-turn dialogue. (2) 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 responses. (3) Professional scholars annotate all the samples with labels, hallucination subtypes 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. The advantages of ours compared with previous benchmarks are listed in Table 1. **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 benchmark. Our contributions can be listed as follows:

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

Table 1: The comparison between our DiaHalu and other benchmarks. 'By LLMs', 'Faith Halu', 'Multi Dia', and 'Explanation' mean whether it is naturally generated by LLMs, whether it provides faithfulness hallucination, whether it is at multi-turn dialogue level, and whether there are explanations respectively (Appendix A.1).

- To our current knowledge, we propose the first dedicated dialogue-level hallucination detection benchmark for LLMs named DiaHalu.

- DiaHalu covers four multi-turn dialogue domains along with five hallucination subtypes extended from factuality and faithfulness hallucination, which are more widely applicable in real-world scenarios.

- The experimental results indicate that DiaHalu is a highly challenging benchmark for most LLMs and existing detection methods, holding significant value for further researches.

2 Related Work

2.1 Hallucination Detection Benchmarks

In recent years, researchers propose numerous benchmarks for hallucination detection.

In earlier years, hallucination detection benchmarks are primarily organized through manual methods or generated via conventional language models. FactCollect (Ribeiro et al., 2022) is an artificially generated, multi-source factual hallucination detection benchmark. Muhlgay et al. (2023) collects error samples by instructing the language model based on pre-defined error types. HADES (Liu et al., 2022) and BEGIN (Dziri et al., 2022c) constitute hallucination detection datasets by conventional language model BERT (Devlin et al., 2019) and T5 (mostly) (Raffel et al., 2020) respectively. These benchmarks are not naturally generated by LLMs as in daily usage.

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

Researchers also raise many sentence-level (Manakul et al., 2023; Zhao et al., 2023c; Guan et al., 2023) and passage-level (Yang et al., 2023; Feng et al., 2023b; Li et al., 2023b) hallucination detection benchmarks. Nevertheless, the dialogue pattern holds broader applications within LLMs. While previous researches rarely, to our current knowledge, propose a dialogue-level hallucination detection benchmark for LLMs. So, our DiaHalu is at the dialogue level.

2.2 Hallucination Detection

Current methods for hallucination detection (Tonmoy et al., 2024) can mainly be divided into four categories. (1) Model-Based. This method involves having the language models perform a classification task to determine whether hallucinated contents are present (Zhao et al., 2023a; Maharaj et al., 2023). (2) Retrieval-Based. For the limited knowledge within the parameters of language models, we can detect hallucination by extracting or retrieving relevant knowledge from external knowledge graphs (Martino et al., 2023; Chen et al., 2023c) or web information sources (Béchard and Ayala, 2024). (3) Sample-Based. Another feasible method is to rewrite the generated contents to evaluate the consistency (Manakul et al., 2023; Zhao et al., 2023c; Zhang et al., 2023a). (4) Proxy-Based. The mainstream view of this method (Zhang et al., 2023b; Gupta et al., 2024) is that 'the lower the probability of generating a token, the more likely a model is to produce hallucination'.

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., 2023a). 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 subtypes 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, shown in Figure 2. Their specific descriptions are in Appendix A.2.

Knowledge-grounded dialogue is designed for users to engage in knowledge-based dialogue with LLMs (Ghazvininejad et al., 2018). The two speakers take part in a conversation about a knowledge-based issue.

Task-oriented dialogue is in a form of human-computer interaction, intending to accomplish a user-specified task (Wen et al., 2017).

Chit-Chat dialogue involves open-ended and non-goal dialogue (Sun et al., 2021). We provide two LLMs with personas and facilitate a dialogue between them.

Reasoning dialogue Following previous works (Chen et al., 2023b; Li et al., 2023a; Buszydlik

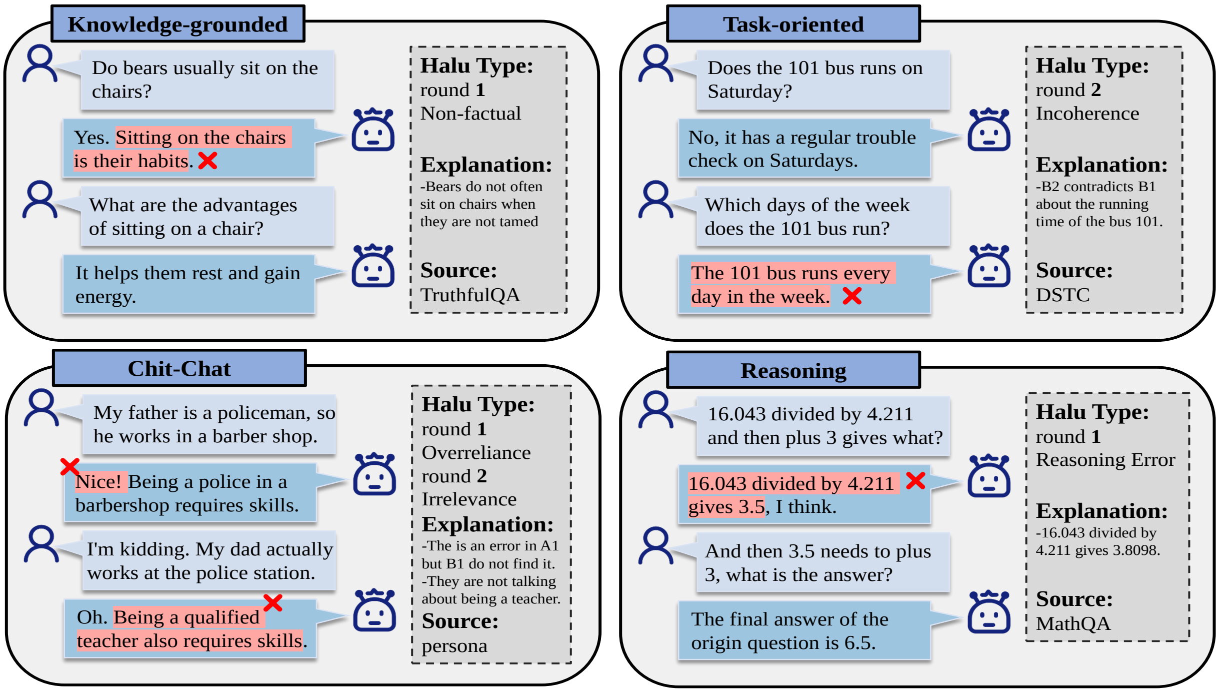

Knowledge-grounded
Do bears usually sit on the chairs?
Yes. Sitting on the chairs is their habits.
What are the advantages of sitting on a chair?
It helps them rest and gain energy.
Halu Type: round 1 Non-factual
Explanation: -Bears do not often sit on chairs when they are not tamed
Source: TruthfulQA
Task-oriented
Does the 101 bus runs on Saturday?
No, it has a regular trouble check on Saturdays.
Which days of the week does the 101 bus run?
The 101 bus runs every day in the week.
Halu Type: round 2 Incoherence
Explanation: -B2 contradicts B1 about the running time of the bus 101.
Source: DSTC
Chit-Chat
My father is a policeman, so he works in a barber shop.
Nice! Being a police in a barbershop requires skills.
I'm kidding. My dad actually works at the police station.
Oh. Being a qualified teacher also requires skills.
Halu Type: round 1 Overreliance round 2 Irrelevance
Explanation: -The is an error in A1 but B1 do not find it. -They are not talking about being a teacher.
Source: persona
Reasoning
16.043 divided by 4.211 and then plus 3 gives what?
16.043 divided by 4.211 gives 3.5, I think.
And then 3.5 needs to plus 3, what is the answer?
The final answer of the origin question is 6.5.
Halu Type: round 1 Reasoning Error
Explanation: -16.043 divided by 4.211 gives 3.8098.
Source: MathQA

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

et al., 2024; Grover et al., 2024; Zheng et al., 2023; Huang and Chang, 2023), we also treat reasoning errors as a kind of hallucination. We have the models discuss mathematical problems to achieve the answers (Kakarla et al., 2024).

3.3 Hallucination Taxonomy

We consider both factuality and faithfulness hallucination (Huang et al., 2023a). Based on Chen et al. (2023b); Wu et al. (2023a); Dziri et al. (2022b) and early works on text coherence (Wolf and Gibson, 2004; Atwell et al., 2024), we carry out a detailed classification of faithfulness hallucination into Incoherence, Irrelevance, and Overreliance. Meanwhile, we introduce Reasoning Error within the reasoning dialogue (Chen et al., 2023b; Li et al., 2023a; Buszydlik et al., 2024; Grover et al., 2024; Zheng et al., 2023; Huang and Chang, 2023). The example of each type can be referenced in Figure 2.

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 trusts in the correctness of the context, generating responses for unanswerable contents (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 from 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 covers 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 them. Additionally, we make use of mathematical problems to assess the logical **reasoning dialogue** abilities 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 Appendix A.3.

4.2 Dialogue Generation

Once finishing collecting the dialogue topics for each domain, we leverage ChatGPT3.5 and GPT4 to generate conversations in the format of self-dialogue. The complete process of dialogue generation is illustrated in Figure 3.

Initially, we integrate the dialogue topics into two system prompts, which are then inputted separately into two LLMs (both are ChatGPT3.5 or GPT4). 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.4. 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 dialogue.

We provide one output sample and more generation details in Appendix A.5. The statistical information of the whole benchmark is in Table 2.

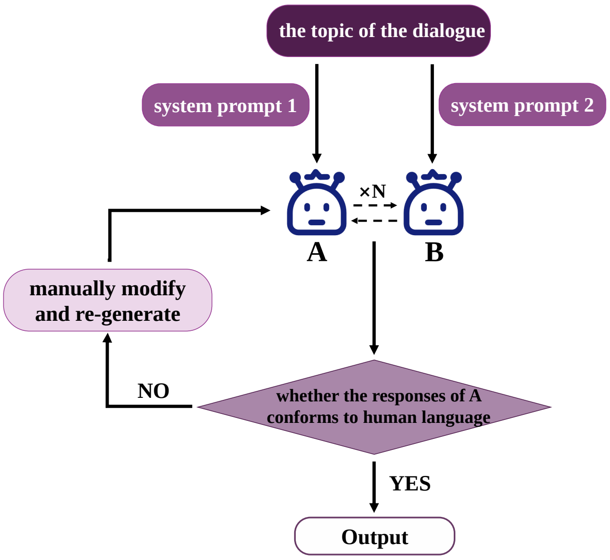

the topic of the dialogue
system prompt 1
system prompt 2
×N
A
B
manually modify and re-generate
NO
whether the responses of A conforms to human language
YES
Output

Figure 3: The complete process of dialogue generation.

Attribute	Attribute Value
Benchmark Name	DiaHalu
Generated by	ChatGPT3.5 / GPT4
Sample Form	dialogue-level
Sample Numbers	1103
Dialogue Rounds	6-10
Avg. Rounds	6.9120
Domain Numbers	4
Hallucination Subtypes	5
Max. Response Length (Words)	183
Avg. Response Length (Words)	13.2899

Table 2: The statistical information of the benchmark.

4.3 Human Annotation

Annotating the hallucination and its types in this dataset is a very challenging task. Since there may be more than one instance of hallucination in multi-turn dialogue. Also, some hallucination subtypes 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 dataset are all seasoned researchers in the field of linguistics and natural language processing. We invite experienced experts in the field of LLMs' hallucination detection from both academia and industry to engage in discussion and conduct sampling checks. For more details about the annotators and the experts, refer to Appendix A.6 (The annotators).

Annotation process is divided into three steps. (1) Each annotator labels some samples for each domain, followed by a careful discussion between the annotators and the experts. The discussions intricately define the application scope of each hallucination label (discussion results in Appendix A.7). (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.6 (The Annotation Process).

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. The final score of Fleiss's Kappa is 0.8842, representing almost perfect agreement among all the annotators. For more calculation details, please refer to Appendix A.6 (Label Consistency).

Annotation Results After annotating the entire dataset, we conduct several statistical analyses on it. Table 3 reveals the probability of hallucination occurring in each dialogue domain. The results indicate that hallucination 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 4 presents the proportion of each hallucination subtype in each dialogue domain. Irrelevance, incoherence, and overreliance widely exist 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 hallucination are defined as errors in reasoning. This statistical information can help us understand the subtypes of hallucination in LLMs' multi-turn dialogue, facilitating an exploration of their origins and contributing to the elimination of these subtypes of hallucination.

	Knowledge	**Task**	**Chit**	**Reasoning**	**Overall**
# Number	371	210	263	259	1103
# Non-Halu	199	135	164	129	627
# Halu	172	75	99	130	476
Halu Rate (%)	46.36	35.71	37.64	50.19	43.16

Table 3: The statistical information of hallucination on the four dialogue domains. '# number', '# Non-Halu','# Halu' and 'Halu Rate' represent the number of samples, the number of samples without hallucination, the number of samples with hallucination and the proportion of hallucinated samples.

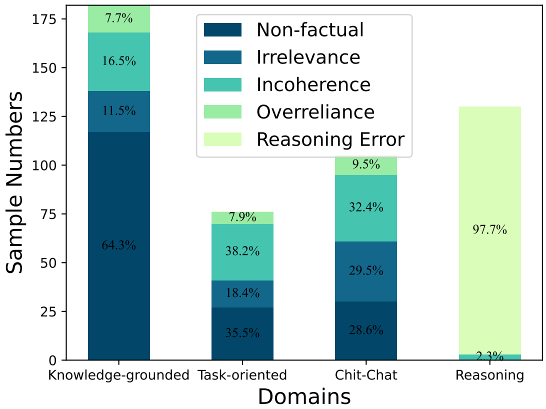


Figure 4: The distribution of five different hallucination subtypes within the four dialogue domains.

5 Experiments

In this section, we assess the performance of several evaluation models and specialized methods on the dataset we introduced. Thereby, we can trial the effectiveness of existing methods in detecting dialogue-level hallucination. We still conduct more fine-grained detection and explore whether the phenomenon of hallucination snowballing exists.

5.1 Baselines

We select some powerful LLMs to detect hallucination by providing specific prompts. These models include open-source LLMs: LLaMa-30B (Touvron et al., 2023), Vicuna-33B (Chiang et al., 2023), and some closed-source LLMs: Gemini1.5 PRO (Anil et al., 2023), ChatGPT3.5 (Wu et al., 2023c) and GPT4 (OpenAI, 2023). Similarly, we also experiment on specialized existing hallucination detection methods, such as FaithCritic (Dziri et al., 2022a), SelfCheckGPT (Manakul et al., 2023) and FOCUS (Zhang et al., 2023b). For a detailed description of the above baselines, please refer to Appendix A.8 (I. Baselines Selected).

5.2 Metrics

For hallucination detection, we use standard binary classification to determine whether there exists hallucination (Table 4). Meanwhile, we also conduct

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	41.57	43.02	42.29	31.86	48.00	38.30	38.46	50.51	43.67	49.61	49.23	49.42	40.72	47.06	43.66
SelfCheckGPT$_B$	42.55	23.26	30.08	35.38	30.67	32.86	30.00	18.18	22.64	60.81	34.61	44.12	43.00	26.47	32.77
SelfCheckGPT$_N$	59.46	25.58	35.77	38.84	62.67	**47.96**	45.19	47.47	46.30	70.58	18.46	29.27	48.65	34.03	40.05
SelfCheckGPT$_P$	55.22	21.51	30.96	48.00	32.00	38.40	45.00	45.45	45.23	62.37	44.62	52.02	52.90	34.45	41.73
FOCUS	46.11	48.26	**47.16**	34.09	60.00	43.48	36.56	49.49	42.06	50.56	34.62	41.10	41.49	46.64	43.92
LLaMa-30B	37.50	5.23	9.18	30.77	5.33	9.09	50.00	11.11	18.18	81.25	10.00	17.81	49.33	7.78	13.43
Vicuna-33B	45.45	5.81	10.31	42.86	4.00	7.32	36.36	4.04	7.27	51.35	14.62	22.75	46.75	7.56	13.02
Gemini1.5 PRO	80.00	20.93	33.18	60.00	36.00	45.00	70.37	38.38	**49.67**	73.63	51.54	60.63	71.49	35.29	47.26
ChatGPT3.5	25.00	0.58	1.14	33.33	2.67	4.93	55.56	5.05	9.26	57.14	6.15	11.11	48.48	3.36	6.27
GPT4	80.89	31.98	45.83	74.19	30.67	43.40	67.74	21.21	32.31	74.07	61.54	**67.23**	75.21	37.61	**50.14**

Table 4: The classification results on four kinds of baselines, and the best F1 scores are in bold form. The indices B, N and P of SelfCheckGPT denote scoring with BERTScore, with NLI and using prompts, respectively.

Method	Knowledge-grounded			Task-oriented			Chit-Chat			Reasoning			Overall		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
FaithCritic w/retrieval	28.26	84.54	42.38	-	-	-	-	-	-	51.63	79.17	62.50	-	-	-
Gemini1.5 PRO	80.00	20.93	33.18	60.00	36.00	45.00	70.37	38.38	49.67	73.63	51.54	60.63	71.49	35.29	47.26
w/ CoT	81.25	22.67	35.45↑	69.77	40.00	50.85↑	75.00	36.36	48.98↓	72.92	53.85	61.95↑	74.47	36.76	49.23↑
w/ retrieval	86.04	21.51	34.42↑	-	-	-	-	-	-	76.70	60.77	67.81↑	-	-	-
ChatGPT3.5	25.00	0.58	1.14	33.33	2.67	4.94	55.56	5.05	9.26	57.14	6.15	11.11	48.48	3.36	6.27
w/ CoT	45.45	2.91	5.46↑	33.33	2.67	4.94	40.00	4.04	7.34↓	47.06	6.15	10.88↑	43.18	3.99	7.31↑
w/ retrieval	70.00	4.01	7.69↑	-	-	-	-	-	-	70.58	9.23	16.32↑	-	-	-
GPT4	80.89	31.98	45.83	74.19	30.67	43.40	67.74	21.21	32.31	74.07	61.54	67.23	75.21	37.61	50.14
w/ CoT	86.05	21.51	34.42↓	73.17	40.00	51.72↑	80.56	29.29	42.96↑	71.43	76.92	74.07↑	75.38	41.18	53.26↑
w/ retrieval	77.89	43.02	55.43↑	-	-	-	-	-	-	74.40	71.54	72.94↑	-	-	-

Table 5: The results of CoT and retrieval techniques on the three closed-source LLMs. The ↑ and ↓ indicate whether CoT and retrieval can promote improvements in F1 score. We also provide the detection results of FaithCritic.

more fine-grained hallucination-type recognition to judge the specific subtype of hallucination and use micro-F1 score for all hallucination categories (Table 6). Appendix A.8 (II. Metrics Calculation) provides more thorough explanations.

5.3 Main Results

From the results in Table 4, we can get the following conclusions.

First, DiaHalu is a highly challenging benchmark for dialogue-level hallucination detection. Except for GPT4, the F1 scores of all other detection methods and detecting LLMs do not exceed 50.00. Existing LLMs, such as LLaMa-30B and Vicuna-33B, are not effective in accurately discerning most samples that involve hallucination. Regarding the specialized detection methods FOCUS and SelfCheckGPT (applying prompt and NLI methods), they achieve relatively better performances. However, it proves challenging with BERTScore for SelfCheckGPT.

Second, ChatGPT3.5 shows 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 hallucinated or not, not along the samples generated by GPT4. So, the majority of its output labels are "Non-Halu".

	NF	Ic	Ir	Ov	RE	ALL
Gemini1.5 PRO	18.97	30.49	11.36	4.76	45.41	26.72
ChatGPT3.5	1.16	4.26	0.00	0.00	9.66	3.93
GPT4	29.38	25.00	5.71	4.65	55.66	32.30

Table 6: Fine-grained hallucination-type recognition F1 scores for three LLMs. ‘NF’, ‘Ic’, ‘Ir’, ‘Ov’ and ‘RE’ stand for Non-factual, Incoherence, Irrelevance, Overreliance and Reasoning Error, respectively. ‘ALL’ represents micro-f1 of all hallucination subtypes.

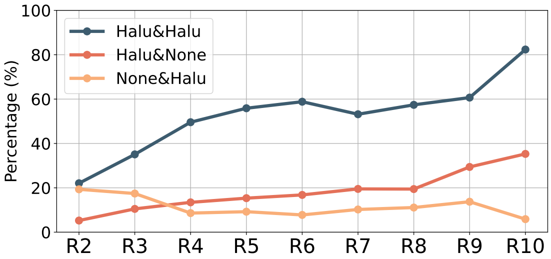


Figure 5: The proportions of the three dialogue round categories. For example, the three values of R7 denote the proportions of ‘these three categories in the 7th round’ within ‘hallucinated samples that have at least seven rounds dialogues’.

Third, the faithfulness hallucination is more difficult to detect for LLMs. Apart from the specialized hallucination detection methods, 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. Nevertheless, task-oriented and chit-chat domains mainly consist of three subtypes of faithfulness hallucination, which requires a LLM to possess long-term memory and the ability to recognize topics/roles transition in dialogue.

5.4 Chain-of-Thought and Retrieval for Detection

Chain-of-Thought (CoT) and Retrieval are two important techniques for enhancing the ability of LLMs. In this section, we test whether these two techniques can improve the effectiveness of hallucination detection in Table 5. More details are depicted in Appendix A.11.

The experimental results indicate that both methods have facilitating effects on hallucination detection. However, Gemini1.5 PRO and ChatGPT3.5 with CoT show a decrease of around 1.00 F1 points in the chit-chat domain. We believe that these two models inherently lack the ability to recognize faithfulness hallucination, and additional CoT contents introduce noise to their judgments.

5.5 Fine-grained Hallucination-type Recognition

Table 6 shows fine-grained hallucination-type recognition results for three open-source LLMs. We can conclude that ChatGPT3.5 fails the recognition of almost all labels. To some extent, Gemini1.5 PRO and GPT4 have the ability to recognize factuality hallucination and reasoning errors, but they have lower F1 scores for the three subtypes of faithfulness hallucination. This result reveals that faithfulness hallucination remains a pressing issue for LLMs.

5.6 Hallucination Snowballing

In this section, we study the hallucination snowballing phenomenon (Zhang et al., 2023c) in our benchmark. Specifically, for each round of dialog (2-10) in all hallucinated samples, we define three categories: **I** hallucination that appears in previous rounds and also appears in the current round (Halu&Halu), **II** hallucination that appears in previous rounds but not appear in the current round (Halu&None) and **III** hallucination that not appear in previous rounds but appears in the current round (None&Halu). We calculate the proportions of these three categories in Figure 5.

First, **I** is greater than the other two categories (**II** and **III**), which means that hallucinated contents are more likely to generate new hallucinated responses. Second, **I** shows the most obvious increasing trend, indicating that the probability of hallucination increases with the number of dialogue rounds. These two findings validate the hallucination snowballing phenomenon.

6 Conclusion

In this paper, we propose a dialogue-level hallucination evaluation benchmark named DiaHalu. We construct the benchmark in a three-step process. The DiaHalu covers four multi-turn dialogue domains and five hallucination subtypes. Experiments through some well-known LLMs and specialized detection methods on the benchmark show that it is a challenging task, holding significant value for further research (Appendix A.12).

Limitations

This paper proposes a novel dedicated dialogue-level hallucination detection evaluation benchmark named DiaHalu. The benchmark covers four multi-turn dialogue domains and five hallucination sub-types. 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 or GPT4 API Keys result in a significant expense. Simultaneously, achieving consistency among all annotators led to prolonged discussion time and money cost. (2) We do not partition the dataset into training, validation, and test sets. The primary purpose of evaluation benchmarks is to assess a models' capabilities. However, if we divide the dataset into the above three categories, this is about assigning capabilities to models. From the perspective of the two objectives, there is a clear difference. Another reason is that we need to consider the black-box detection scenario for those closed-source LLMs. However, if a division into these three types of datasets is necessarily required, it would require more data samples and larger resource consumption.

Ethics Statement

The benchmark is primarily generated by ChatGPT3.5 or GPT4. We obtain all the API Keys through a paid subscription. All the annotators are real people 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. *Preprint*, arXiv:1905.13319.

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.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

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

Katherine Atwell, Mert Inan, Anthony B. Sicilia, and Malihe Alikhani. 2024. Combining discourse coherence with large language models for more inclusive, equitable, and robust task-oriented dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 3538–3552. ELRA and ICCL.

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

Patrice Béchard and Orlando Marquez Ayala. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. *CoRR*, abs/2404.08189.

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.

Aleksander Buszydlik, Karol Dobiczek, Michal Teodor Okon, Konrad Skublicki, Philip Lippmann, and Jie Yang. 2024. Red teaming for large language models at scale: Tackling hallucinations on mathematics tasks. *CoRR*, abs/2401.00290.

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.

Kedi Chen, Jie Zhou, Qin Chen, Shunyu Liu, and Liang He. 2024. A regularization-based transfer learning method for information extraction via instructed graph decoder. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1472–1485. ELRA and ICCL.

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.

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

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.

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, Sivan Milton, Mo Yu, Osmar R. Zaïane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? 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 5271–5285. Association for Computational Linguistics.

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

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.

Shresth Grover, Vibhav Vineet, and Yogesh S. Rawat. 2024. Navigating hallucinations for reasoning of unintentional activities. *CoRR*, abs/2402.19405.

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.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. *CoRR*, abs/2404.10136.

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.

Michael Heck, Nurul Lubis, Benjamin Matthias Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, Carel van Niekerk, and Milica Gasic. 2023. Chatgpt for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 936–950. Association for Computational Linguistics.

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.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. 2023a. Deceiving semantic shortcuts on reasoning chains: How far can models go without hallucination? *CoRR*, abs/2311.09702.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. 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. 2024a. Leveraging large language models for nlg evaluation: A survey. *Preprint*, arXiv:2401.07103.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024b. Leveraging large language models for NLG evaluation: A survey. *CoRR*, abs/2401.07103.

Jiangyi Lin, Yaxin Fan, Feng Jiang, Xiaomin Chu, and Peifeng Li. 2023. Topic shift detection in chinese dialogues: Corpus and benchmark. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part III*, volume 14189 of *Lecture Notes in Computer Science*, pages 166–183. Springer.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. 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.

Kishan Maharaj, Ashita Saxena, Raja Kumar, Abhijit Mishra, and Pushpak Bhattacharyya. 2023. Eyes show the way: Modelling gaze behaviour for hallucination detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11424–11438. Association for Computational Linguistics.

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.

Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (LLM) hallucination. In *The Semantic Web: ESWC 2023 Satellite Events - Hersonissos, Crete, Greece, May 28 - June 1, 2023, Proceedings*, volume 13998 of *Lecture Notes in Computer Science*, pages 182–185. Springer.

Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. 2023. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. 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 14408–14416. AAAI Press.

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. *Preprint*, arXiv:2305.15852.

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. *Preprint*, arXiv:2303.08774.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*, abs/2401.01313.

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. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *CoRR*, abs/2310.07521.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is chatgpt a good NLG evaluator? A preliminary study. *CoRR*, abs/2303.04048.

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.

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.

Florian Wolf and Edward Gibson. 2004. Representing discourse coherence: A corpus-based analysis. In *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*.

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023a. A multi-task dataset for assessing discourse coherence in chinese essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6673–6688. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. 2023b. 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. 2023c. 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.

Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. *CoRR*, abs/2403.18243.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Kumar Sricharan. 2023a. Sac^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *CoRR*, abs/2311.01740.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 915–932. 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. 2023c. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Lingjun Zhao, Khanh Nguyen, and Hal Daumé III. 2023a. Hallucination detection for grounded instruction generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4044–4053. Association for Computational Linguistics.

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. 2023b. 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. 2023c. 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? *Preprint*, arXiv:2304.10513.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press.

A Appendices

A.1 The Comparison with Other Benchmarks

In Table 1, 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. (2023b), Manakul et al. (2023) and Yang et al. (2023) respectively.

A.2 The Four Dialogue Domains

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 dialogue (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.

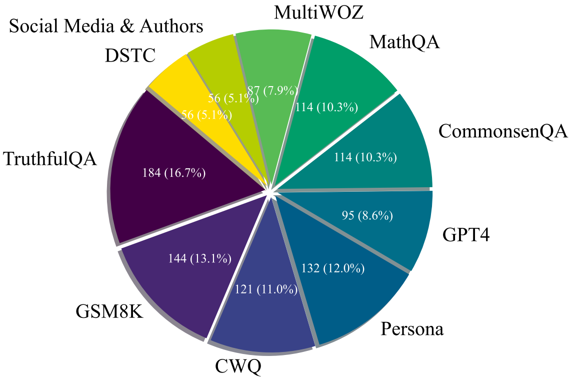


Figure 6: The distribution of the 10 sources for dialogue topics.

Reasoning dialogue centralizes on the logical reasoning and understanding capabilities of LLMs. Following previous works (Chen et al., 2023b; Li et al., 2023a; Buszydlik et al., 2024; Grover et al., 2024; Zheng et al., 2023; Huang and Chang, 2023), we also treat reasoning errors as a kind of hallucination. We have the models discuss mathematical problems to achieve the answers (Kakarla et al., 2024).

A.3 The Distribution of the Sources for Dialogue Topics

The number and proportion of all dialogue topics in DiaHalu across the 10 topic sources are shown in Figure 6.

A.4 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)[1] and GPT4 (gpt-4-1106-preview)[2] 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 7, Figure 8, Figure 9 and Figure 10.

Previous works prove the LLMs' ability to follow complex instructions (Manakul et al., 2023; Guan et al., 2023; Chen et al., 2023b; Mündler et al., 2023; Li et al., 2023b; Wang et al., 2023b; Liu et al., 2023; Li et al., 2024b), including some hallucination tasks. Thus, we reference such kinds of prompts, and then we formulate prompts for our benchmark.

---

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo
[2] https://openai.com/gpt-4

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 7: The brief system prompts for knowledge-grounded dialogue.

A.5 Generation Details and Data Format

We utilize 8 ChatGPT3.5 and 2 GPT4 API Keys, generating approximately 4000 examples in total initially. We conduct a thorough examination of samples to filter the confused formats or contents not aligned with instructions (system prompts). To ensure a balanced number of positive and negative samples, we manually remove samples with highly similar semantics. Finally, the number of samples generated by ChatGPT3.5 and GPT4 are 748 and 355 respectively. The total number is 1103. The reason we use GPT4 to generate data is to ensure its competitiveness for a long time in the future.

Previous well-known hallucination benchmarks (Manakul et al., 2023; Guan et al., 2023; Chen et al., 2023b; Yang et al., 2023) contain 238, 400, 847, and 300 samples respectively, which is at the similar scale as ours. What's more, our benchmark is at dialogue level, which contains about 7620 rounds (more than 27600 rounds initially) of interactions in total (6.9120 average rounds per dialogue as shown in Table 2). This indicates a larger volume of data compared to previous benchmarks at the sentence and passage levels.

It is also worth noting that, given that we assume the two subjects of the dialogue are A and B, both A and B are set to be either ChatGPT3.5 or GPT4, and it is not possible for one to be ChatGPT3.5 and the other to be GPT4.

We provide the specific format of one sample from the benchmark in Figure 11.

A.6 The Supplementary Details for Annotation

The annotators all obtain at least a bachelor's degree, get a high score in IELTS or TOEFL exams, and are proficient in using search engines such as Google and Bing. The annotators are all seasoned researchers in the field of linguistics and natural language processing. In addition, everyone exhibits strong collaborative and communicative skills. We also invite senior experts in the field of hallucination detection from academia and industry to engage in discussions and data checking.

The Annotation Process To ensure the annotation quality, we perform three steps for annotation as described in Section 4.3 (Annotation process). **First**, each annotator labels around 50 samples for each domain. The annotators are required to label the presence of hallucination, hallucination subtypes and locations, along with the corresponding explanations. For cases of inconsistent annotation, we invite experts to provide suggestions in a discussion. After that, annotators specify the appli-

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 8: The brief system prompts for task-oriented dialogue.

cation scope of each hallucination label as needed. Then the annotators take a vote for resolving the label-inconsistency of the first 50 samples in each domain. **Second**, the entire dataset is annotated according to this standard. The annotators label all the rest samples and vote for the inconsistent samples, following data checks by the experts. **Third**, we conduct data statistics of the whole dataset.

In the first step above, the discussion is organized in the form of online meetings. Annotators provide the inconsistent-labeled samples to experts (first 50 in each domain), after which all annotators and experts agree on a time for an online meeting discussion. Experts provide suggestions, and annotators modify the application scopes of hallucination labels based on the suggestions, thus making it more reliable.

The Price The annotation time for each sample ranges from 2 to 10 (average 6.2) minutes. We pay each annotator 0.5 US dollars for annotating a sample and pay each expert 0.5 US dollars for checking a sample. This exceeds the local average hourly wage. Through the aforementioned approach, the quality of the annotations and the value of the benchmark are ensured. Thus, we consider it is greatly contributory to propose such a benchmark.

Label Consistency 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:

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

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 9: The brief system prompts for chit-chat dialogue.

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

$$Fleiss'sKappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

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 0.8842, representing almost perfect agreement among all the annotators.

The Label Platform We use Label Studio[3] for labeling, which is an online open-source data labeling platform in the field of artificial intelligence. The annotation interface is depicted in Figure 12.

A.7 The Application Scope of Hallucination Labels

Annotating the hallucination and its subtypes in this benchmark is a very challenging task. One of the reasons is that some hallucination subtypes in edge cases are defiant to differentiate. Therefore, in the first stage of annotation, we provide detailed definitions for each hallucination subtype. Below are the results of the discussion between the experts and the annotators.

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

---

[3] https://labelstud.io/

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.

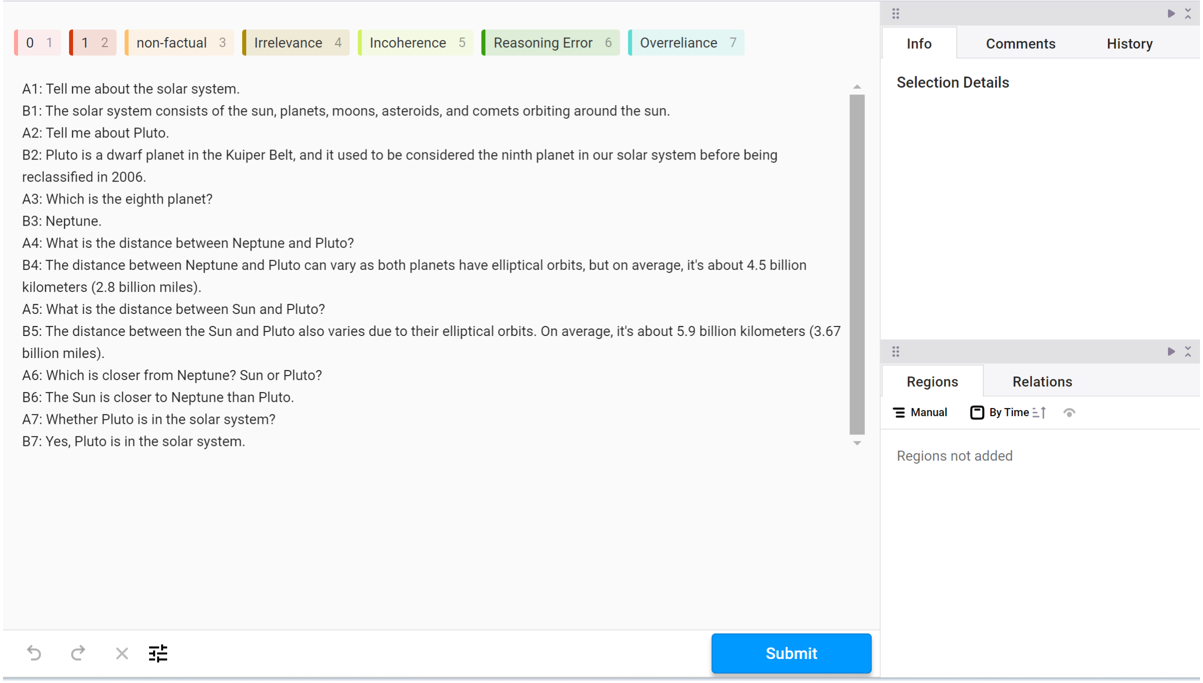
Figure 10: The brief system prompts for reasoning dialogue.

{
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",
Which LLM: "ChatGPT3.5"
}

Figure 11: One sample from our benchmark DiaHalu

Incoherence means there is a contradiction between one of the speakers and himself in the dialogue context, a contradiction between the two speakers (when both are declarative sentences) in the dialogue context. It also refers to factual and relevant nonsense, contextually inappropriate responses or other inconsistent errors.

Irrelevance involves responses that are irrelevant to the dialogue topic or an irrelevance due to misunderstanding the grammar of a question. (Please note that we emphasize the use cases for both interrogative and declarative sentences.)



A1: Tell me about the solar system.
B1: The solar system consists of the sun, planets, moons, asteroids, and comets orbiting around the sun.
A2: Tell me about Pluto.
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.
A3: Which is the eighth planet?
B3: Neptune.
A4: What is the distance between Neptune and Pluto?
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).
A5: What is the distance between Sun and Pluto?
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).
A6: Which is closer from Neptune? Sun or Pluto?
B6: The Sun is closer to Neptune than Pluto.
A7: Whether Pluto is in the solar system?
B7: Yes, Pluto is in the solar system.

Figure 12: The annotation interface on Label Studio.

Overreliance is that the LLM excessively trust in the correctness of the context, generating serious responses to statements that were inherently wrong or unanswerable (in a declarative sentence).

Reasoning Error covers all errors within the reasoning dialogue.

A.8 Baselines and Metrics

I. Baselines Selected

Below is a detailed description of all the baselines we selected.

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

FaithCritic (Dziri et al., 2022a) is one of the most effective dialog text hallucination classifiers before the era of large language models. Trained on a large-scale dialog corpus, it can output the confidence level for each classification label. Since the model's input includes dialog-related knowledge, we use the retrieved contents as the knowledge during the experiment.

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.

FOCUS (Zhang et al., 2023b) is an improved version of SelfCheckGPT. It takes into account the attention scores between entity tokens, enabling more accurate classification of hallucination at both the sentence and paragraph levels.

LLaMa-30B && Vicuna-33B They are two well-pretrained and widely deployed open-source LLM backbones[4] (Touvron et al., 2023; Chiang et al., 2023). We provide a specially designed prompt to assist with detection. More details about this prompt are shown in Appendix A.9.

Gemini1.5 PRO Gemini1.5 PRO[5] (Anil et al., 2023) is the latest version of the language model launched by Google. It inherits the powerful natural language processing capabilities of its predecessor and has made significant improvements in understanding and generating text. We employ the same prompt for binary detection as LLaMa-30B

---

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

[5] https://gemini.google.com/

and Vicuna-33B do. We also create a manually prompt to assist with fine-grained recognition in Appendix A.10.

ChatGPT3.5 && GPT4 Both of these models are developed by OpenAI[6]. ChatGPT3.5 marks the beginning of the era of large language models and GPT4 is currently the most powerful language model (Wu et al., 2023c; OpenAI, 2023). We employ the same prompt as LLaMa-30B and Vicuna-33B do for binary detection. And the same prompt as Gemeni1.5 PRO do is used for fine-grained recognition. The ChatGPT3.5 version is ChatGPT3.5 (gpt-3.5-turbo)[7] and the GPT4 version is GPT4 (gpt-4-turbo) [8].

II. Metrics Calculation

Despite we annotating the subtypes of hallucination in the dataset, achieving consistent labels even among humans requires further discussion. Therefore, similar to past hallucination detection efforts, we first focus on a binary classification task of determining the existence of hallucination. Consequently, we utilize binary classification evaluation metrics: Precission, Recall and F1. The positive label for this classification task is set as "Halu", for our main focus is testing the model's ability to recognize hallucination.

As for more fine-grained hallucination-type recognition, We define a correct judgment as one where both the presence of hallucination and the specific subtype of hallucination are accurately identified. For all label types, we use micro-F1 score to quantify the performances of the three classification models.

A.9 The Prompt Designed for Detection

In Figure 13, we show the whole prompt specially designed for hallucination detection of the baselines: LLaMa-30B, Vicuna-33B, Gemini1.5 PRO, ChatGPT3.5, and GPT4. 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.

Previous works prove the LLMs' ability to follow complex instructions (Manakul et al., 2023; Guan et al., 2023; Chen et al., 2023b; Mündler et al., 2023; Li et al., 2023b; Wang et al., 2023b; Liu et al., 2023; Li et al., 2024b), including some hallucination tasks. Thus, we reference those prompts that classify hallucination using LLMs, and then we formulate prompts for ours.

A.10 The Prompt Designed for Fine-grained Recognition

In Figure 14, we show the whole prompt manually created for fine-grained hallucination-type recognition of the three closed-source baselines: Gemini1.5 PRO, ChatGPT3.5, and GPT4.

A.11 The Settings for CoT and Retrieval

Chain-of-Thought (CoT) (Feng et al., 2023a) describes the organized sequence of logical reasoning that unfolds during thinking. Retrieval (Gao et al., 2023; Ye et al., 2024) means retrieving relevant contents from the media to supplement external knowledge for LLMs. We employ CoT in all four domains to enhance the performance of ChatGPT3.5 and GPT4. The specific CoT is illustrated in Figure 15. Even so, only knowledge-grounded and reasoning domains are tested with retrieval via Google[9]. 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.12 Future Works

The necessity of dialogue-level The sentence-level, passage-level and dialogue-level hallucination differ in the hallucination types and detection difficulties. We will explain this with the following examples to make it more clear.

Here is a sentence-level hallucination example from the dataset FactCHD (Chen et al., 2023c). 'User: Can you tell me which mountain range is longer, the Alps or the Pyrenees? LLMs: The Pyrenees are longer than the Alps.' In this sentence-level example, the LLM only responds to the user's question with one sentence. We need to assess whether there are hallucination in the single sentence generated by the LLM.

Here is a passage-level hallucination example from the dataset WikiBio (Manakul et al., 2023). "Matthew Aylmer, 1st Baron Aylmer was an Irish soldier and colonial administrator. He was born in Dublin, the son of a barrister, and was educated at Trinity College, Dublin. ... He was buried in Westminster Abbey." This passage-level example is directly generated by the LLM. We need to determine whether the passage with multiple sentences

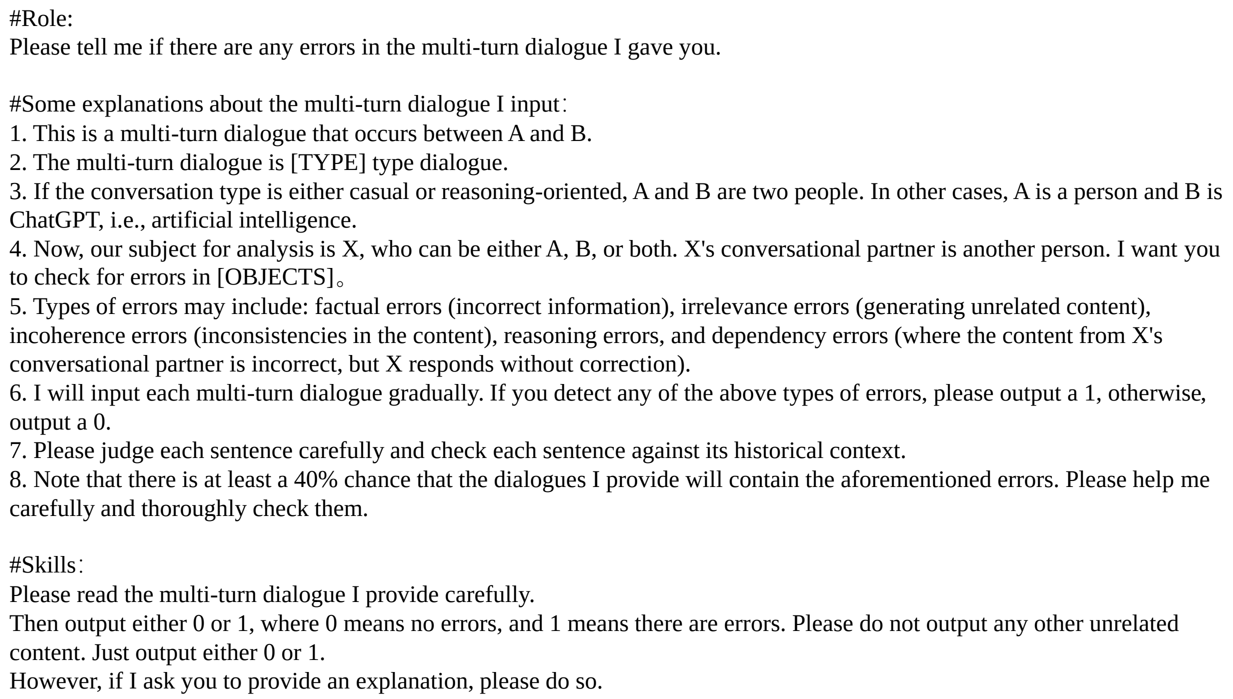
---

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

[7] https://platform.openai.com/docs/models/gpt-3-5-turbo

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

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

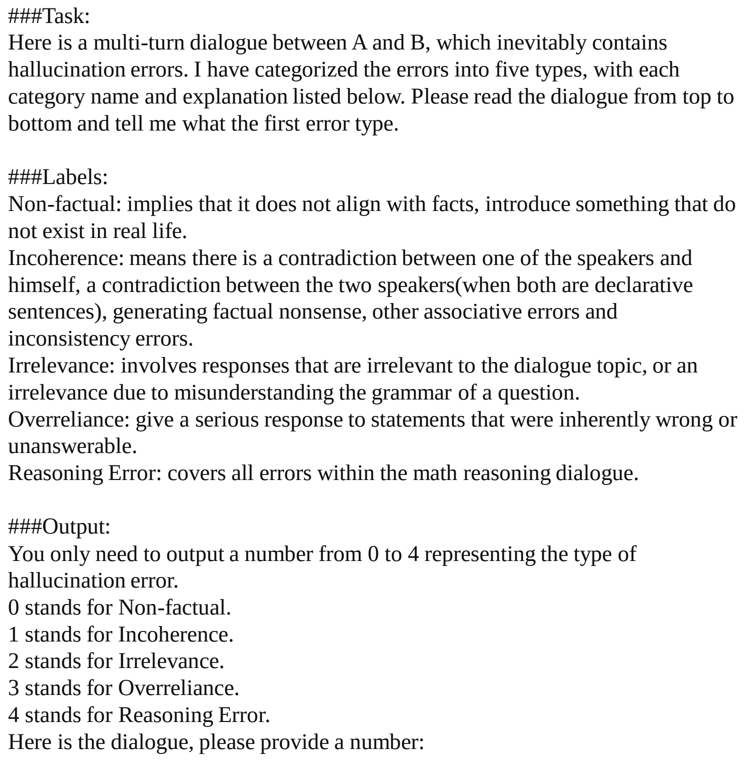


#Role:
Please tell me if there are any errors in the multi-turn dialogue I gave you.

#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 13: The whole prompt for hallucination detection of the baselines.



###Task:
Here is a multi-turn dialogue between A and B, which inevitably contains hallucination errors. I have categorized the errors into five types, with each category name and explanation listed below. Please read the dialogue from top to bottom and tell me what the first error type.

###Labels:
Non-factual: implies that it does not align with facts, introduce something that do not exist in real life.
Incoherence: means there is a contradiction between one of the speakers and himself, a contradiction between the two speakers(when both are declarative sentences), generating factual nonsense, other associative errors and inconsistency errors.
Irrelevance: involves responses that are irrelevant to the dialogue topic, or an irrelevance due to misunderstanding the grammar of a question.
Overreliance: give a serious response to statements that were inherently wrong or unanswerable.
Reasoning Error: covers all errors within the math reasoning dialogue.

###Output:
You only need to output a number from 0 to 4 representing the type of hallucination error.
0 stands for Non-factual.
1 stands for Incoherence.
2 stands for Irrelevance.
3 stands for Overreliance.
4 stands for Reasoning Error.
Here is the dialogue, please provide a number:

Figure 14: The whole prompt for fine-grained hallucination-type recognition.

involves hallucination. In this example passage, hallucination occurs in the last sentence which provides unfactual information. Since there are interrelations or dependencies between the sentences in the passage, passage-level hallucination detection is more challenging than the sentence-level one.

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 15: The whole CoT for the four domains of dialogue.

The examples of dialogue-level hallucination are shown in Figure 2, which covers four domains and five hallucination types. The differences between it with sentence-level and passage-level hallucination are as follows: First, more types of hallucination occur frequently in dialogue. One possible reason for most benchmarks merely focusing on detecting factuality hallucination is that they are organized at a sentence or passage level. In this setting, faithfulness hallucinatios (including Incoherence, Irrelevance, and Overreliance) are less likely to occur. In contrast, since dialogue generation requires LLMs to have context coherence (Mishra et al., 2023), track the dialogue state (Heck et al., 2023), possess long-term memory capabilities (Zhong et al., 2024), and have the ability to recognize topic shifts (Lin et al., 2023), faithfulness hallucination (including Incoherence, Irrelevance, and Overreliance as described in lines 261-263) occur more frequently in dialogue. In our benchmark, the faithfulness hallucination mainly accounts for Task-oriented dialogue and Chit-Chat dialogue as shown in Figure 4. **Second**, it is more challenging to detect hallucination from dialogue than a single sentence or a passage. Since a dialogue contains multiple rounds of interactions that are interdependent, it can not determine whether the current round has hallucination merely based on the current content. For example, the Incoherence hallucination type occurs as the answer is not consistent with the previous context in the task-oriented dialogue in Figure 2. Thus, it needs to analyze the context dependency in the whole dialogue and judge coherence, relevance and reasoning correctness, spanning multiple rounds of interactions for dialogue-level hallucination detection, which is more challenging than sentence-level and passage-level detection.

Therefore, it is necessary to construct a dialogue-level hallucination evaluation benchmark to promote research in LLM studies.

Dialogue-level hallucination detection is an important work in the future. We propose the first dedicated dialogue-level hallucination detection evaluation benchmark for LLMs and experimental results show that it is a very challenging task. Therefore, combining previous works (Chen et al., 2017; Deriu et al., 2021), 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 sentence-level or passage-level factuality hallucination (Luo et al., 2024). Hallucination elimination at the dialogue level not only requires models to have much parameter knowledge, but also a long-context memory capabilities, the abilities to recognize changes in topics/roles and logical transitions in the dialogue. These are helpful in addressing faithfulness hallucination. At the same time, improving the accuracy of knowledge in the LLMs' parameters and the reasoning abilities are equally important.

Hallucination snowballing is the phenomenon that LLMs tend to accumulate hallucination rather than self-correcting during the generation process (Zhang et al., 2023c). Some previous works validate this phenomenon (Azaria and Mitchell, 2023; Ang et al., 2023). In our benchamark, there is a noticeable issue of hallucination snowballing. Through the experimental results, we also display such phenomenon. Eliminating hallucination snowballing in LLMs is extremely urgent 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 and try to find a solution. However, this issue in the application of human-machine interaction and multi-agent scenarios still remains crucial.