Exploring the Factual Consistency in Dialogue Comprehension of Large Language Models

Shuaijie She , Shujian Huang*, Xingyun Wang , Yanke Zhou, Jiajun Chen
National Key Laboratory for Novel Software Technology, Nanjing University
shesj@smail.nju.edu.cn, huangsj@nju.edu.cn
{wangxingyun,zhouyk}@smail.nju.edu.cn, chenjj@nju.edu.cn

Abstract

LLMs (Large Language Models) usually interact with users in the form of dialogue and generate responses following their instructions, which naturally require dialogue comprehension abilities. However, dialogue comprehension is a general language ability which is hard to be evaluated directly. In this work, we propose to perform the evaluation focusing on the factual consistency issue with the help of the dialogue summarization task. Besides evaluating and analyzing the dialogue summarization performance (DIAC-Sum) of different LLMs, we also derive factual questions from the generated summaries and use them as a more flexible measurement of dialogue comprehension (DIAC-FactQA). Our evaluation shows that, on average, 26.8% of the summaries generated by LLMs contain factual inconsistency. Even ChatGPT, the strongest model evaluated, has such errors in 16% of its summaries. For answering the factual questions, which is more challenging, the average error rate of all evaluated LLMs is 36.1%. Both results indicate serious deficiencies. Detailed analysis shows that the understanding of subject/object of the conversation is still challenging for LLMs. Furthermore, to stimulate and enhance the dialogue comprehension ability of LLMs, we propose a fine-tuning paradigm with auto-constructed multi-task data, which achieved a relative error rate reduction of 11% on DIAC-FactQA.[1]

1 Introduction

With the development of large language models (LLMs), such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023a), etc., it has becoming a promising way to interact with users through conversations, where LLMs generate responses following users' instructions in the dialogue. This form of conversational communication naturally requires high dialogue comprehension ability to capture the factual information, which has becoming a prerequisite for successfully completing tasks in the conversation.

Previous studies have evaluated the dialogue comprehension ability of relatively small language models using dialogue question answering (QA) data, e.g., DREAM(Sun et al., 2019) and FriendsQA(Yang and Choi, 2019). However, the questions in these data come from either human exams or random sampling of the dialogue content, neither of which targets on the **DIA**logue **C**omprehension (**DIAC**) ability of LLMs.

To fill this gap, we propose to perform the evaluation with the help of the dialogue summarization task (DIAC-Sum), since summarization extracts important information from the dialogue which naturally requires a correct understanding of the dialogue. The summaries generated by 5 popular LLMs are collected and manually evaluated for factual consistency. The results are reported as an empirical study of the current situation.

More importantly, we derive a set of factual questions from the factual inconsistencies in the generated summaries (DIAC-FactQA), which corresponds to the difficult parts of understanding these dialogues. By answering these questions, the dialogue comprehension ability of LLMs could be evaluated in a more flexible and precise way. Please note that the resulting dataset could be used for evaluating the dialogue comprehension ability of other LLMs in the future, without any more human annotation.

Our evaluation results show that existing LLMs still have serious deficiencies in conversational understanding. On average, 26.8% of the summaries generated by LLMs contain factual inconsistencies. Even the summaries generated by ChatGPT, the strongest LLMs in our evaluation, has 16% identified to be incorrect. Further factual question probing results shows that the average error rate of

* Corresponding author
[1]We will release all data public to facilitate future research

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6087–6100
June 16-21, 2024

Dialogue:		Marsha: Guys, we've planned the trip with John last night as we promised. Cynthia: great, thank you for that. Marsha: but of course you have to agree on that. Mohammad: sure, but I really trust you. Gavin: me too. ...
DIAC-Sum	**Erroneous Summary:**	Marsha, Cynthia, Mohammad and Gavin are going to Madagascar.
	Error: **Error Type:**	The participation of John is neglected. SubObjE
	Corrected Summary:	Marsha, Cynthia, Mohammad, Gavin and John are going to Madagascar.
DIAC-FactQA	**Factual Question 1:** Correct Answer: Wrong Answer:	Who is going to Madagascar? Marsha, Cynthia, Mohammad, Gavin and John Marsha, Cynthia, Mohammad and Gavin
	Factual Question 2: Correct Answer: Wrong Answer:	How many people are going to Madagascar? 5 4
	Factual Question 3: Correct Answer: Wrong Answer:	Is John going to Madagascar? yes no

Table 1: This is an example of our annotation process. The Dialogue is a partial input dialogue from the SAMSUM dataset. The Error Summary is the output of fine-tuned the BART-Large model on this dataset which has SubObjE (missing Subject John). The Corrected Summary is the summary that the annotator has revised with minimal changes. Factual Question 1-3 are constructed by the annotator to test the dialogue comprehension capability.

all evaluated LLMs reaches 36.1%. For the reference, the error rate of ChatGPT and Vicuna-13B is 26.2% and 40.5%. Detailed analysis on different error categories show that the understanding of subject-object is one of the most challenging problem.

We also present attempts to stimulate the dialogue comprehension ability by fine-tuning the model with auto-constructed multi-task data. Experimental results demonstrate that after fine-tuning, the model exhibits a better dialogue comprehension ability, providing a potential direction for future work and improvements.

2 Dialogue Comprehension Benchmark (DIAC)

2.1 Data and Model Preparation

Our evaluation utilized the SAMSum (Gliwa et al., 2019) dataset, which is composed of messenger-like conversations together with their summaries. These conversations cover various topics in real-life messenger chats, ranging from informal to formal and showcasing the use of slang words, emoticons, and typos. This format bears resemblance to the current interaction between users and LLMs. The summaries succinctly summarize the overall conversation in third person. We use the same subset consisting of 150 conversations from the test set of SAMSum as Wang et al., 2022, so our analysis could also be compared with their results.

Our evaluation is based on the generations of the LLMs, and five popular LLMs are selected to initialize the evaluation: Alpaca-7B[2], Vicuna-7/13B[3], Flan T5-11B[4], and ChatGPT[5]. These models vary in size, training methods, and training datasets, but all of them have achieved impressive results and are widely used currently. Latest LLMs such as Mistral (Jiang et al., 2023), LLaMA2 (Touvron et al., 2023b) and GPT4 can also be incorporated into our evaluation in an annotation-free style.

2.2 DIAC-Sum: Inconsistency Annotation and Correction

As summarization requires a comprehensive understanding of the information presented in the conversation, inconsistencies in the summary may indicate incorrect comprehension of the dialogue.

We ask the initial five LLMs to generate summaries for each conversation, resulting 750 summaries. For the purpose of evaluating factual consistency, we require the LLMs to generate summaries that contains exact information in the conversation rather than some vague description. This

[2]https://github.com/tatsu-lab/stanford_alpaca
[3]https://github.com/lm-sys/FastChat#vicuna-weights
[4]https://huggingface.co/google/flan-t5-xxl
[5]https://platform.openai.com/docs/models/gpt-3-5

Settings	DIAC-FactQA	DREAM
Dialogue Length (tokens)	106.0	53.8
Dialogue Turns	12.7	4.4
Speaker Num	2.5	2.0
Question Length (tokens)	8.5	6.9
Questions per dialogue	10.0	1.2
Dialogue Distinct 1-gram	0.17	0.08
Dialogue Distinct 2-gram	0.67	0.42

Table 2: We compile some basic characteristics of the dataset and compare them with DREAM. Distinct (Li et al., 2016) is a common metric used to measure the diversity of a text by the repetition of n-grams; the higher it is, the better the diversity of the text.

is achieved by a sampling and filtering strategy. Please refer to the Appendix A.1 for details.

We manually annotate and correct the factual inconsistencies in the collected summaries according to Wang et al., 2022, where the inconsistencies are categorized into the following five types.

- **SubObjE**: additions, deletions, or substitutions of participants (as subjects or objects) described in the summary.
- **ProE**: incorrect references of pronouns in the summary.
- **HalE**: events in the summary not exist in the dialogue, resembling a hallucination.
- **ParE**: particular errors in the time and location of events.
- **NegE**: contradictions in the summary with events in the original dialogue.

In the manual annotation process, each annotator is assigned with a dialogue and a corresponding summary generated by LLMs. The annotation task is to verify the consistency between the generated summary and the dialogue, and to identify the error along with its type, if present. The annotator is also asked to correct the error summary with minimum editing efforts. The annotation process involves 5 annotators with an agreement of coefficient $k = 0.83$. All 750 summaries are annotated. The performance of each LLMs are evaluated by the factual errors in their respective generated summaries. More details can refer to Appendix A.2.

An example of the annotation of DIAC-Sum is shown in Table 1. Given the conversation, the error summary neglects John's participation in the trip. According to the definition, this inconsistency is classified as SubObjE and the participation of John is integrated into the summary for correction.

2.3 DIAC-FactQA: Factual QA Construction

The result of previous annotation presents the ability of current LLMs, however, the annotation process is expensive and cannot be reused directly. Therefore, we present DIAC-FactQA, which transfer the evaluation of summaries into a question-answering process. Different from previous practice that also using QA to evaluate the dialogue comprehension abilities (Sun et al., 2019; Yang and Choi, 2019), our DIAC-FactQA focuses on inconsistencies in the auto-generated summaries, so it is more accurate in identifying the weaknesses in the model's understanding of the dialogue.

More specifically, we ask the annotators to compare the corrected summary with the error summary and wrote several questions based on the previously identified errors. For each question, two choices of answers are collected, including answers from both the correct summary and the error summary. This process yields a correct answer and a distractor associated with each question, with the dialogue serving as a reference.

In order to ensure comprehensive testing of the model's dialogue comprehension, the annotators are required to create multiple types of questions for each error, as shown in Table 1. In this example, the model overlooks the fact that John is also participating in the trip. To test whether the model actually knows the fact, we annotate three diverse questions from different perspectives, including a question for listing all the participants, a question for the number of participants and a direct question for John's participation.

Besides DIAC-Sum we construct, Wang et al., 2022 also collected 750 summaries, including the result generated by four BART-based models and the original reference summary in SAMSum. For better diversity, we also corrected these summaries and merged two datasets to obtain 1,500 summaries for our DIAC-FactQA annotation, resulting in a total of 446 triples of (dialogue, erroneous summary, corrected summary). In total, we collected 1484 questions together with two options, answer and distractor, based on the inconsistent summaries. On average, an inconsistent summary has 3.32 questions.

Table 2 shows some basic information about our dataset. DIAC-FactQA has higher distinct n-gram, offering better diversity compared to DREAM. Moreover, DIAC-FactQA includes longer dialogues and more turns per conversation, increas-

Model	SubObjE	ProE	HalE	ParE	NegE	Overall
Alpaca-7B	16.0	4.0	18.7	12.7	1.3	40.0
Vicuna-7B	14.0	7.3	12.0	10.7	5.3	34.7
FlanT5-11B	9.3	5.3	2.7	2.0	4.0	18.0
Vicuna-13B	10.7	5.3	7.3	4.0	4.0	25.3
ChatGPT	4.7	5.3	2.7	4.7	0.7	16.0
GPT4	4.0	3.3	3.3	1.3	0.7	11.3
Average	9.8	5.1	7.8	5.9	2.7	24.2
Human-Ref*	2.7	3.3	0.7	2.0	1.3	9.3
BART-based*	14.7	6.0	5.3	14.7	6.0	36.7

Table 3: The proportion of inconsistencie (%) of DIAC-Sum. The row of Average is the average of all 5 LLMs. The rows of Human-Ref and BART-based models come from Wang et al., 2022, for comparison purposes.

Model	SubObjE	ProE	HalE	ParE	NegE	Overall	DREAM
Alpaca-7B	51.6	47.5	58.0	47.7	46.6	48.5	31.0
Vicuna-7B	47.0	46.5	49.0	46.4	48.3	46.3	32.5
FlanT5-11B	25.5	18.8	26.4	30.2	27.5	24.8	4.80
Vicuna-13B	41.6	39.0	42.0	43.4	32.8	40.5	17.9
ChatGPT	25.2	24.6	29.8	28.1	31.4	26.1	6.10
LLaMA2-7B-Chat	48.1	48.0	51.5	50.6	45.1	47.7	29.1
Mistral-7B-Instruct	38.7	37.0	42.0	36.2	31.7	36.7	14.5
GPT4	15.2	14.0	24.2	20.1	22.2	18.5	3.32
Average	36.6	34.4	40.4	37.8	35.7	36.1	17.4

Table 4: Error rate (%) of answering factual questions on DIAC-FactQA and DREAM.

ing its complexity. With a larger set of factual questions, DIAC-FactQA allows for a more comprehensive assessment of dialogue understanding consistency. More details about our dataset can refer to Appendix A.3.

3 Evaluation Results and Analysis

3.1 DIAC-Sum Results and Analysis

Table 3 shows the factual consistency of LLMs on the summarization task. "BRAT-based" are the results of BART-based models after supervised training for summarization; "Human-Ref" refers to the summary written by human. These two results are adopted from Wang et al., 2022. The main observations are as follows.

Inconsistency still plague LLMs: ChatGPT has the best performance, with the overall error rate and each individual error categories being better than those of other models. However, surprisingly, 16% of the summaries still contain inconsistency. This indicates that even on the top-performing ChatGPT, there are still many errors in generation. It is worth noting that FlanT5 was trained on the SAMSum dataset and thus has better performance. In addition, Alpaca and Vicuna were both trained on the instruction datasets without seen this summarization data directly, with error rates ranging from 25.3% to 40%. The average inconsistency rate in the summaries generated by all the models is 26.8%. The problem of factual inconsistency remains a serious issue, which also points to the existing weakness in the comprehension of dialogues by LLMs.

Stronger language abilities reduce errors: Compared with the previous BART models, the LLMs make fewer errors. For example, FlanT5, which is also trained on SAMSum, makes far fewer errors than the BART model, from 36.7% to 16.8%. We also observed results that are similar to other evaluation benchmark (Zheng et al., 2023a; Beeching et al., 2023), where Vicuna 7B showed better performance than Alpaca 7B. Moreover, with parameter scaling up, Vicuna 13B demonstrated even better results. The model's better performance and dialogue comprehension ability are related.

SubObjE and HalE remain the major issues: Similar to BART, SubObjE remains challenging for LLMs. The SubObjE involves understanding the referring information in the dialogue, analyzing the speaker's intention (such as whether to attend an event), handling abbreviations in the dialogue, and other issues. Another noteworthy observation is that the hallucination in the LLMs has increased significantly. How to make the large-scale model

Task Re-Assignment	(SubObjE)
Maya: Bring home the clothes that are hanging outside Boris: I'm not home right now Boris: I'll tell Brian to take care of that	Who will bring home the clothes? ChatGPT: B: Boris Correct: A: Brian
Intention of Participation	(SubObjE)
Marsha: Guys, we've planned the trip with John last night as we promised Cynthia: great, thank you for that Marsha: but of course you have to agree on that Mohammad: sure, but I really trust you Gavin: me to	How many people are going to Madagascar? ChatGPT: B: Four Correct: A: Five
Complex Pronoun References	(SubObjE, ProE)
Ian: I don't know any Claire... Maddie: Really? I spoke with Leah and she told me that you dumped her years ago Ian: are you sure she was talking about me? believe me none was named Claire	Who did Ian dump years ago? ChatGPT: B: Leah Correct: A: Claire
Subsequent Clarification	(SubObjE, HalE)
Timmy: What about food? Gemma: Others and I will cover it Timmy: Others? I thought it was a date :P Gemma: U remember I have a bf, right?	Who will cover the food at the BBQ? ChatGPT: B: Gemma Correct: A: Gemma and her boyfriend
Inference Intent from Expressions	(SubObjE, NegE and HalE)
Brian: lets NOT do the homework and see what happens Lena: do what you want I won't risk it	Is Lena not going to do the homework? ChatGPT: A: Yes Correct: B: No

Table 5: Analysis of problematic dialogue comprehension cases from ChatGPT.

aware of the factual consistency and generate reliable output based on the input dialogue is a problem that requires further attention.

3.2 DIAC-FactQA Result and Analysis

The results of different LLMs on DIAC-FactQA are shown in Table 4. Besides the initial 5 LLMs used for generating summaries, we also extend our evaluation to later LLMs such as LLaMA2, Mistral and GPT4. This extension does not require any further annotation.

Many LLMs have severe deficiencies in dialog understanding: The experiment results on DIAC-FactQA confirm the deficiencies in dialogue understanding of LLMs. Even ChatGPT and GPT4 have error rates of 26.1% and 18.5%, respectively. In addition, Vicuna-13B, which is considered a relatively strong LLMs, has an error rate of 40.5% on these questions. The problem still remains severe on latest models such as LLaMA2, Mistral, and GPT4. Especially for GPT4, which has extraordinary performance currently, there are still 18.5% of questions that cannot be answered correctly. The average error of LLMs even reaches 36.1%, revealing their limited understanding of conversations.

DIAC-FactQA has diagnosed more issues: Compared with DREAM, DIAC-FactQA can more effectively discover these conversation understanding defects. Large models can answer many test questions in the DREAM dataset. For example, Mistral, ChatGPT, and GPT4 has a low error rate of 14.5%, 6.1%, and 3.3%. However, in DIAC-FactQA, their performance are much worse: 36.7%, 26.1%, and 18.5% respectively. DIAC-FactQA contributes to a more comprehensive diagnosis of deficiencies in dialogue understanding and facilitates future improvements.

3.3 Case Study

Previously, we analyze the dialogue comprehension ability of LLMs across inconsistent categories. We also conducted an examination of specific scenarios in which models are susceptible to errors (Table 5).

In daily conversations, there are often task assignments, activity arrangements, etc. When the individuals indicate that they are unavailable and have made **task re-assignments**, LLMs may not fully comprehend the information and neglect the

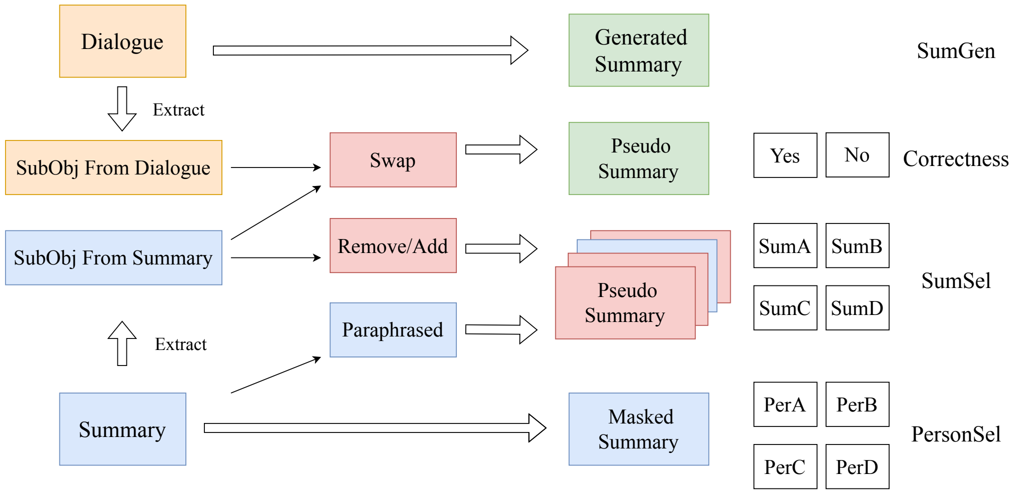

Figure 1: Illustration of our multi-task pseudo-data. The negative examples, highlighted in red, represent constructed instances that contain subject-object errors. Conversely, the positive examples are indicated in blue.

most recent updates. This may result in ChatGPT mistakenly presuming that Boris will complete the task of gathering clothing.

The **intention of participation** corresponds to activity organization commonly found in the dialogues. The model needs to integrate scattered information from the dialogue in order to correctly understand which people are participating in the activities. In this example, the model easily overlooks John and Cynthia.

One notable characteristic of dialogues is the extensive use of pronouns to facilitate communication. The example involves two speakers and two individuals in the conversation, and include a significant number of **complex pronouns**. While humans can comprehend the dialogue, it poses a considerable challenge for the model. The answers of factual question show that the model's understanding of these pronouns are quite confusing.

Meanwhile, temporary uncertainties often arise during discussions, but these are typically resolved through **subsequent clarification** in the dialogue. In the given example, it is explained later that "Others" refers to the boy friend. However, it requires the casual language model to associate the later explanation with the earlier uncertainty, which is also a non-trivial challenge.

Inference intention in dialogue presents another challenge. Speakers often use a variety of expressions to convey their intentions, such as the phrase "I'm not one to take risks" in the example. It is crucial for the model to accurately interpret whether the user intends to express agreement or disagreement with the various stances presented.

4 Multi-Task Fine Tuning Paradigm

4.1 Pseudo Dataset Creation

The above analysis reveals that the primary deficiency of current LLMs in conversation understanding is the recognition of subject and object in dialogue. Therefore, we made some initial attempts to improve the dialogue understanding ability of LLMs by focusing on their subject and object comprehension capabilities.

Our intuition is simple: the model must correctly understand the subject-object in the dialogue to perform the tasks accurately. More important, in order to drive and validate the model's genuine learning of general dialogue understanding abilities, rather than simply fitting the distribution and task forms, the synthesized data we created differs from the evaluation data regarding its distribution, task structures, and instructions. Our data construction is based on the SAMSum dataset. Figure 1 is an overview of our construction process. In addition, Table 6 provides an example of data construction. Further details, including the instructions we designed for each task, will be discussed in the Appendix A.5.

Summary is a brief overview of the core content in a conversation. To understand the dialogue, a clear comprehension of the subject and object in its summary is essential. Therefore, we propose the **PersonSel** (Person Selection) by masking the person in the reference summary, and requiring the model to select the correct option from four choices based on the dialogue to fill in the masked position. Moreover, dialogue summarization is a comprehen-

Input Dialogue	Jenkin: hey what is your spirit animal? Sophie: what? Sophie: I dont know, a fox lol Jenkin: are you wiley? Sophie: sometimes Jenkin: I am a dolphin ...
Candidate Summary	...Sophie would choose a fox.
SumGen	Jenkin have be read about spirit animals and he be draw to a dolphin. Sophie would choose a fox.
Correctness	The candidate summary is inconsistent or not? **Label:** A: No
SumSel	Identify the correct summary of the dialogue from the provided options A: Jenkin have read about spirit animals and he be draw to a dolphin. Sophie would choose a fox. Jenkin and wiley will bring pack of cards with spirit animals to Sophie tomorrow. ... D: Jenkin have read about spirit animals and he draw to a dolphin. Sophie would choose a fox. Jenkin will bring pack of cards with spirit animals to Sophie tomorrow. **Label:** D
PersonSel	Based on the conversation, select the most suitable option to fill in the [MASK] blank. ...[MASK] would choose a fox. A: Jenkin B: Sophie C: Sophie and Jenkin D: Jenkin **Label:** B: Sophie

Table 6: An example of multi-task pseudo data

sive form of learning dialogue comprehension. We also sampled some dialogue summarization data for training as **SumGen** (Summary Generation).

Aside from training models to deduce correct answers, teaching models to recognize errors is also a viable approach. Models require the ability to comprehend dialogues and infer the correct subject-object in order to accurately identify errors. To obtaining negative examples, we extract subjects and objects from the dialogue and summary respectively, including speakers and person entities[6]. For the summary's subject-object, we simulate SubObjE by substitution, addition, and deletion to create negative examples. Meanwhile, we can generate positive examples using a paraphrasing model.

Once we have obtained positive and negative examples, we can construct **Correctness** (Correctness Decision) data to teach model which is correct or not. We further simplified this task to **SumSel** (Summary Selection) by providing four options for the model to choose. One of these is correct, while the remaining three contain subject-object errors. As a re-ranking task, it enables the model to learn the required abilities in an easier way.

4.2 Experimental Settings

Based on the training set of the SAMSUM dataset, we automatically sampled 10,000 data for each task,

	DIAC-FactQA		DREAM
Model	**SubObjE**	**Overall**	**Overall**
Alpaca-7B	44.1(-7.95%)	42.9(-11.5%)	21.2(-31.6%)
Vicuna-7B	42.9(-8.72%)	42.0(-9.30%)	22.6(-30.5%)
Vicuna-13B	34.4(-17.3%)	35.6(-12.1%)	15.2(-15.1%)
Average	40.5(-11.3%)	40.2(-11.0%)	19.6(-27.6%)

Table 7: Question answering error rate (%) on DIAC-FactQA and DREAM after multi-task fine-tuning. Error reduction rate is caculated relatively.

with a total of 40,000. At the same time, we hope to enhance the ability of dialogue comprehension while maintaining the original ability of the model as much as possible. Therefore, we mixed pseudo-data and Alpaca's instruction-following data. For efficient fine-tuning, we used LoRA (Hu et al., 2021) to fine-tune the model. Each model was trained for 3 epochs and the training process took only 1 hour on 8 A100 80G. We run three times of each model with different random seeds and report the average results.

4.3 Fine Tuning Result and Analysis

Table 7 shows the performance of different LLMs after fine-tuning. Our data not only enhances the model's understanding of subject/object, but also improves the overall dialogue comprehension ability. It has achieved stable improvements on two datasets, which are 11% and 27.6% respectively.

[6] https://spacy.io/models/en#en_core_web_lg

Settings	DIAC-FactQA	DREAM
Baseline	48.5	31.0
All	42.9	21.2
w/o SumGen	45.3	24.7
w/o Correctness	46.2	25.4
w/o SumSel	43.9	22.3
w/o PersonSel	46.0	28.5

Table 8: Overall question answering error rate (%) on DIAC-FactQA and DREAM in ablation study.

As we designed, there exist notable difference between our training data and test data which highlights the efficacy of our data in enhancing the model's core ability of dialogue understanding, which can generalize across different tasks. Although LLMs have overcome more difficulties in dialogue understanding with the help of our data, it can be observed that there are still many errors. DIAC-FactQA remains a challenging tasks that requires further in-depth research in the future.

There is a reasonable gap between the performance in Table 7 and FlanT5 in Table 3. We believe that this is mainly due to FlanT5 has been fully fine-tuned on a larger and more diverse training data set (including the SAMSum and DREAM and other QA data in a format similar to test set).

4.4 Ablation Study

In order to validate the impact of each task on the learning process, we remove one task at a time and fine-tune Alpaca 7B using the remaining data. The experimental result in Table 8 demonstrates that PersonSel had the most significant impact, and the result is consistent over two test sets. This finding suggests that teaching the LLMs to reason about the identity of the person in the summary could effectively enhance the model's ability to comprehend the conversation.

Moreover, Correctness and SumGen also play a crucial role. On the one hand, for the model to decide whether a summary of the dialogue is correct, it should be capable of comprehending the information in the conversation accurately. On the other hand, as previously discussed, summarizing a dialogue involves understanding the dialogue, especially the salient information in it. Finally, even though SumSel shows some impact, it is relatively limited. We assume that this might be because SumSel has lengthy inputs, comprising a question, a dialogue and four summary, which could be challenging for the model to learn.

5 Related Work

5.1 Dialogue Summarization and Question Answering

Dialogue summarization is an important task in natural language generation (Tuggener et al., 2021), which involves summarizing the main content of a dialogue into a third-person summary. It can be used for various scenarios such as meetings (McCowan et al., 2005; Zhong et al., 2021), customer service (Liu et al., 2019), and daily chats (Gliwa et al., 2019). Due to the dispersed information in multi-turn utterances, abstractive summarization is naturally suitable for modeling dialogue summarization (Wang et al., 2022). There are some works that have proposed dialogue-based question answering datasets, such as DREAM (Sun et al., 2019) which is derived from listening test data, and some (Yang and Choi, 2019; Ma et al., 2018) are question-answering data based on dialogue data such as TV drama scripts and constructed question by manual and rule-based methods.

5.2 Faithfulness in Dialogue Summarization

Recent research has focused on factual consistency in dialogue summarization. Tang et al. proposed CONFIT, which identified and annotated several error types. Meanwhile, Huang et al. used this data to evaluate the effectiveness of BARTScore through unlikelihood training on negative samples. Wang et al. identified shortcomings in CONFIT's methodology, scope, and analysis, re-identifying six types of inconsistencies. They annotated and analyzed the faithfulness of four BART fine-tuned dialogue summarization models and found that 35% of the generated summaries contained errors.

5.3 Investigating Large Language Models

In recent years, large language models represented by GPT-3 (Brown et al., 2020) have demonstrated impressive language capabilities across various NLP tasks, such as sentence classification, question answering, machine translation, among others. Some studies have fine-tuned large models on task-specific data to enable them to interact with humans and respond to their instructions effectively; such models include FlanT5 (Chung et al., 2022) and ChatGPT. Alpaca (Taori et al., 2023) and Vicuna (Chiang et al., 2023) collected interaction data from large model APIs to construct the instruction dataset, and they trained impressive instruction-following models. While these models are flourishing, some research has explored the task capabilities of large models. For example, Qin et al. (2023) conducted a simple comparison of 20 datasets across 7 NLP tasks. However, only ChatGPT was included and only ROUGE (Lin, 2004) metrics were compared. In contrast, the aim of this study is to provide a deeper understanding of the dialogue comprehension capability.

6 Conclusion

In this paper, we conducted a detailed evaluation and analysis of the factual consistency in dialogue comprehension of existing LLMs, and attempted to improve upon them. Specifically, we annotated the consistency of summaries generated by the LLMs, and constructed factual questions from generated errors to evaluate the existing models. The evaluation results showed that there are still serious dialogue comprehension defects in current LLMs. Therefore, we made some initial attempt by using automatically constructed multi-task pseudo data to fine-tune the model. The experimental results showed that after fine-tuning, the model's dialogue comprehension capabilities were indeed enhanced, providing a possible solution for further research.

7 Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02), research project of Nanjing University-China Mobile Joint Institute.

Limitations

Our data is manually annotated, which inevitably introduces certain biases. However, we have employed double annotation confirmation to minimize these biases as much as possible. Our study primarily focuses on the dialogue understanding abilities of LLMs, rather than covering all aspects. This means that our conclusions may not fully represent the performance of LLMs in other tasks or domains. We need to be cautious about its misuse and also need more analysis of the abilities of LLMs in the future. Meanwhile, due to resource constraints, this study mainly evaluated relatively popular large models and was unable to cover all existing large models. Finally, as a preliminary attempt to improve general dialogue understanding ability, the data construction strategy we used is relatively simple. Further in-depth research is needed to improve the dialogue understanding capabilities of LLMs in the future.

Ethics Statement

The authors have no conflicts of interest. The datasets used come from publicly available sources and are compliant with their published license. And all data used and created in this paper does not include sensitive content such as personal information in real world. LLMs were only used to generate summaries and answer factual questions, in accordance with their intended use.

References

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

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

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. CoRR, abs/2210.11416.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. CoRR, abs/1911.12237.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv: 2106.09685.

Sicong Huang, Asli Celikyilmaz, and Haoran Li. 2022. Ed-faith: Evaluating dialogue summarization on faithfulness. arXiv preprint arXiv:2211.08464.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. arXiv preprint arXiv: 2310.06825.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Chunyi Liu, Peng Wang, Jiang Xu, Zang Li, and Jieping Ye. 2019. Automatic dialogue summary generation for customer service. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1957–1965.

Kaixin Ma, Tomasz Jurczyk, and Jinho D. Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In North American Chapter of the Association for Computational Linguistics.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, volume 88, page 100. Citeseer.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? CoRR, abs/2302.06476.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. Transactions of the Association for Computational Linguistics, 7:217–231.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Confit: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. NAACL.

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

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288.

Don Tuggener, Margot Mieskes, Jan Milan Deriu, and Mark Cieliebak. 2021. Are we summarizing the right way?: a survey of dialogue summarization data sets. In Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic (online), 7-11 November 2021, pages 107–118. Association for Computational Linguistics.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 4897–4908. Association for Computational Linguistics.

Zhengzhe Yang and Jinho D. Choi. 2019. Friendsqa: Open-domain question answering on tv show transcripts. In SIGDIAL Conferences.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pretrained transformer language models. arXiv preprint arXiv: 2205.01068.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, et al. 2021. Qmsum: A new benchmark for query-based multidomain meeting summarization. arXiv preprint arXiv:2104.05938.

A Appendix

A.1 Filtering Rules for Summary Generation

When collecting summaries from LLMs, we noticed that some models generate results that do not contain exact information of the conversation, such as using the vague term "the group" to refer to people, and using words such as "the dialogue" to start a vague summary. These summaries cannot reflect the level of the model's understanding of the conversation, and are not suitable for the evaluation of factual consistency. Therefore, if the model produces a result that includes any of the phrases listed in Table 9, we reject the result and prompt the model to generate new one.

A.2 Annotation Details

We recruited five annotators and explicitly informed them that the purpose of the annotated data is to evaluate the conversation comprehension ability of large language models. We fistly presented and explained each error types defined in Wang et al., 2022 and its examples for each annotator. Then we instructed the annotators to annotate using the following guidelines: "Read the dialogue carefully. Determine whether the summary contains factual inconsistency based on the dialogue. If so, identify the category of error and correct it. For each error, write three questions with as much variety as possible. Use the corrected summary to answer the questions as the correct option, and the erroneous summary to answer the questions as the distractor." During the entire annotation process, the payment for the annotators was sufficient. The use and public access of the data has been approved, and we will make the data publicly available after the end of the anonymous period to facilitate future research.

Phrases
"In this dialogue", "In the dialogue", "In this conversation", "In the conversation","The conversation revolves", "The dialogue", "The speaker","The group", "A group","The conversation"

Table 9: When the generated summary contains one of these phrases, we will discard the results and ask the model to generate again.

Type	E-Num	Q-Num	Q-Num/E-Num
SubObjE	184	709	3.86
ProE	82	282	3.41
HalE	99	364	3.68
ParE	127	441	3.47
NegE	59	204	3.45

Table 10: We report the number distribution of inconsistency error (E-Num) and our constructed factual question (Q-Num) over inconsistency types

A.3 Distribution of Error and Factual Question

We performed a detailed statistical analysis to examine the distribution of errors and the associated factual questions across various categories. Our findings, as presented in Table 10, indicate a parallel trend between the distribution of questions and errors, with SubObjE emerging as the most common and NegE as the least. Upon assessing the mean number of questions elicited per error, it is noted that the disparity in the mean quantity of questions among different error categories is comparatively minimal. Notably, SubObjE consistently elicits the highest mean number of questions, whereas ProE is the lowest.

A.4 General Ability is Preserved while Improving the Consistency

In our study, we take the Alpaca-7B model as an example to investigate the impact of our fine-tuning strategy on on the general capability, while enhancing the consistency of dialogue understanding. To this end, we evaluated the model on the MTBench benchmark (Zheng et al., 2023b) in Single-Score Mode and Pairwise Mode. In Single-Score Mode, the generated answers will be rated by GPT-4 on a scale from 1 to 10. In Pairwise Mode, GPT-4 will be presented with two answers from different models at the same time and required to select the superior one. The experimental results indicate that the single score is increased from 4.86 to 5.02 after fine-tuning. Meanwhile, the pairwise comparisons show that the fine-tuned model produced better answers in 36.3% of the cases, while only 33.7% for the original model (with a 30% tie rate). These findings suggest that our method not only effectively enhances the model's consistency in dialogue understanding but also successfully retains its general capability, demonstrating the efficacy and versatility of our approach.

A.5 Instruction for Each Task

Table 11 shows a subset of instruction templates that were used by us for training and evaluation purposes. It can be observed that the training data instructions and the test instructions used have significant differences for different tasks, in both content and form. While the Correctness instruction requires answering yes or no questions, PersonSel and SumSel are four-option multiple-choice questions, and SumGen requires generating a summary. This design aims to enable the model to learn core conversational understanding abilities and avoid fitting shallow task formats only.

Task	Instruction
DIAC-Sum:	Please summarize the following conversation as brief as possible.
DIAC-FactQA: **DREAM:**	Here is a dialogue, a question, and two options below. Please read the dialogue carefully and select the best option to answer the question. Your answer should be either A or B.
QE:	1. Please confirm whether the summary accurately reflects the participants involved in the event based on the conversation. 2. Reviewing the summary, did anyone participate in the event is inconsistent with the dialogue? Yes or no answer please. 3. According to the dialogue, are any of the person in the summary incorrect or missing? Please respond with yes or no. ...
PersonSel	1. Based on the conversation, select the most suitable option to fill in the [MASK] blank. 2. After considering the discussion, which of the four choices would be the best to complete the [MASK] position? 3. What is the most fitting option to complete the [MASK] blank based on the conversation? ...
SumSel	1. There is a dialogue below and four candidate summaries, of which three contain errors. Please select the correct one. 2. Identify the correct summary of the dialogue from the provided options. 3. From the given choices, choose the summary that correctly represents the dialogue without any errors. ...
SumGen	1. Write a brief summary of this conversation 2. Based on the dialogue, Summarize the main points 3. Please provide a concise summary of the conversation ...

Table 11: Instructions used for generating the summarization and perform question-answering. We used the same instructions for tests in DIAC-FactQA and DREAM. For each training task, we randomly selected three instructions from a total of ten to present here.