FineDialFact: A benchmark for Fine-grained Dialogue Fact Verification

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are known to produce hallucinations—factually incorrect or fabricated information—which poses significant challenges for many Natural Language Processing (NLP) applications, such as dialogue systems. As a result, detecting hallucinations has become a critical area of research. Current approaches to hallucination detection in dialogue systems primarily focus on verifying the factual consistency of generated responses. However, these responses often contain a mix of accurate, inaccurate or unverifiable facts, making a factual label overly simplistic and coarse-grained. In this paper, we introduce a benchmark, FineDialFact, for fine-grained dialogue fact verification, which involves verifying atomic facts extracted from dialogue responses. To support this, we construct a dataset based on publicly available dialogue datasets and evaluate it using various baseline methods. Experimental results demonstrate that methods incorporating Chain-of-Thought (CoT) reasoning can enhance performance in dialogue fact verification. Despite this, the best F1-score achieved on the HybriDialogue, an open-domain dialogue dataset, is only 0.748, indicating that the benchmark remains a challenging task for future research. Our dataset and code will be public on GitHub.

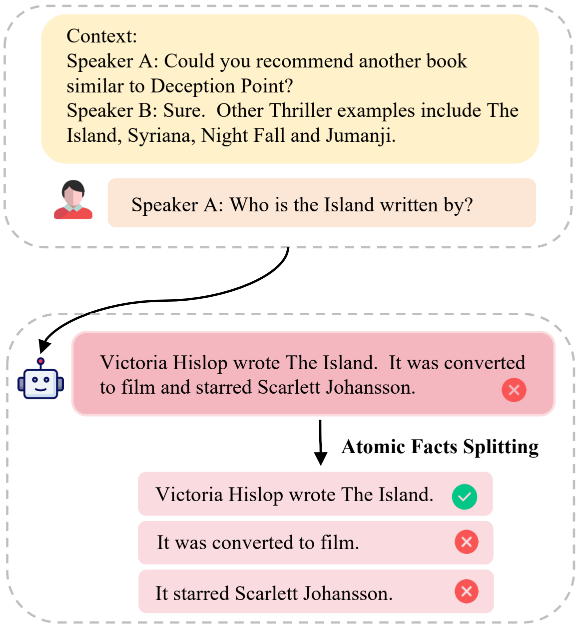

Figure 1: An example of the response-level based dialogue fact verification and fine-grained one. The difference between them is that the latter is based on the atomic facts split by a response.

1 Introduction

In recent years, large language models (LLMs) have demonstrated impressive capabilities across a wide range of tasks. However, one persistent challenge is hallucination — the generation of factually incorrect or misleading content. This issue is particularly concerning in dialogue systems, where hallucinated responses can mislead users and potentially pose risks to social trust and stability.

Previous approaches to hallucination detection on dialogue systems mainly focus on human evaluation (Ni et al., 2023; Li et al., 2022; Shuster et al., 2021; Yu et al., 2022), while evaluating hallucination manually is time-consuming and labour-intensive. Several automatic methods have been proposed to detect hallucinations, including uncertainty estimation (Farquhar et al., 2024) and fact verification (Chen et al., 2024). Uncertainty estimation has two main limitations: (1) it relies on token-level probabilities, which are often inaccessible or unreliable in closed-source models; and (2) if a model is overly confident in hallucinated content, the estimation can be misleading. Therefore, our work focuses on direct fact verification to address these issues. Existing fact verification on dialogue systems (Chen et al., 2024; Gupta et al., 2021) verify the responses based on the external knowledge and dialogue, and then output the label: “Supports”, “Refutes” and “Not Enough Information”. However, these methods only focus on the response level and output a factual label, ignoring that the response may contain true, hallucinated or non-verifiable facts. As shown in the Figure 1, labeling the response as entirely incorrect is overly coarse, as it includes at least one accurate fact.

To address the above limitation, we systematically research the fine-grained fact verification for dialogue systems and offer a benchmark, named FineDialFact. We first split the dialogue response into small pieces of sentences, called atomic facts. Each atomic fact is then verified independently using external knowledge and large language models (LLMs). As there are no existing available datasets, we construct a dataset manually from two public dialogue datasets, OpendialKG and HybriDialogue, and report their inter-agreement by Cohen's Kappa. To evaluate the dataset, we provide a set of metrics: accuracy, precision, recall, F1-score and Cohen's Kappa for measuring the performance in different perspectives.

In addition, we evaluate a series of Chain-of-Thought (CoT) based approaches using Flan-T5, Llama, and GPT models on our constructed datasets. These CoT approaches include zero-shot CoT, few-shot CoT, and CoT distillation. For zero-shot CoT, we simply add a reasoning prompt such as "Let's think step by step." For few-shot CoT, we manually annotate a set of samples and use GPT-4o to generate corresponding reasoning steps, then retrieve the top-N most relevant samples as demonstrations. For CoT distillation, we use GPT-4o to annotate data and generate CoT reasoning processes, which are then used to fine-tune smaller language models. The experimental results show that CoT series-based methods are able to improve the performance of LLMs significantly. However, dialogue fact verification on the HybriDialogue dataset remains a challenging task, with the highest F1-score reaching only 0.748, achieved by GPT-4o.

The contributions can be listed as follows:

1. We delve into the fine-grained fact verification for dialogue systems, named FineDialFact. To the best of our knowledge, this is the first systematic research on the automatic fine-grained factuality evaluation of dialogue systems.
2. We provide a newly constructed dataset for evaluating the fine-grained dialogue factuality, laying a foundation for further research in this area.
3. We evaluate several baselines, including CoT series approaches. The experimental results show that the HybriDialogue dataset is more challenging, and the highest score, achieved by GPT-4o, is only 0.748, opening up new challenges for future research.

2 Related Work

2.1 Dialogue Hallucination Detection

Hallucination in LLMs has attracted increasing attention in open-domain dialogue systems. While current hallucination detection approaches often rely on human evaluation (Ni et al., 2023; Yu et al., 2022), this method is time-consuming and labor-intensive, highlighting the need for effective automatic evaluation methods. Recent advances in automatic dialogue hallucination detection have shown promising progress. Honovich et al. (2021) introduced a metric that leverages question generation and answering with natural language inference to evaluate factual consistency without reference responses. Li et al. (2023a) developed a comprehensive benchmark of 35,000 samples that not only enabled detection but also demonstrated that external knowledge and intermediate reasoning steps can significantly improve performance. Building on factuality detection, Mishra et al. (2024) proposed a more nuanced approach by categorizing hallucinations as Entity, Relation, Sentence, Invented, Subjective, and Unverifiable types. Most recently, Chen et al. (2024) addressed a critical gap with DiaHalu, the first dedicated dialogue-level hallucination benchmark that extends beyond factuality to include faithfulness hallucinations (incoherence, irrelevance, and overreliance) across multiple dialogue domains.

Their work revealed the concern that inaccuracies accumulate during multi-turn dialogues rather than being self-corrected. Despite these advances, current methods still struggle with the complex nature of dialogue responses, where a single response may consist of a long sentence containing both verifiable and non-verifiable parts. To address this, we split such responses into multiple short, atomic facts, allowing each to be evaluated separately.

2.2 Fine-grained Detection

Recent progress in assessing the quality of open-source dialogue generated by LLMs has led to a variety of methodological improvements. Early studies primarily used traditional similarity-based metrics, such as n-gram overlap, and later incorporated embedding-based methods with pre-trained models to better capture semantic similarity between generated and reference texts (Zhang et al., 2019). However, these single-score metrics often fall short in capturing the full complexity of long-form or dialogic outputs. As a result, recent research has turned toward more fine-grained evaluation approaches, especially within natural language processing and dialogue systems. Inspired by tasks like aspect-based sentiment analysis—which acknowledges the possibility of having both positive and negative sentiments in the same sentence (Tan et al., 2019)—researchers have started exploring span-level or unit-level evaluation methods. For instance, Song et al. (2024) and Wan et al. (2024) introduced fine-grained techniques for detecting hallucinations in text summarization, allowing for more accurate identification of factual errors. Likewise, Min et al. (2023) proposed a fine-grained fact scoring method to evaluate factual accuracy in long-form text generation, although its use has so far been limited to bio-generation and does not directly apply to dialogue. Similarly, Zhong et al. (2022) developed a unified, multidimensional evaluation framework for text generation tasks to support fine-grained analysis. Building on these efforts, our work seeks to combine fine-grained evaluation with hallucination detection to more effectively assess the quality and factual consistency of open-source dialogue systems.

2.3 Chain of Thought

Since the emergence of LLMs, there has been increasing interest in applying their capabilities to various downstream tasks in NLP. One notable development is the CoT approach, introduced by Wei et al. (2022), which involves a sequence of intermediate reasoning steps designed to enhance LLM performance—particularly for complex tasks such as math word problems, commonsense reasoning, and symbolic computation. To address the challenge of requiring hand-crafted examples for few-shot CoT prompting, Zhang et al. (2022) proposed an automated method for collecting examples by clustering similar samples. Additionally, Kojima et al. (2022) demonstrated that LLMs can perform surprisingly well in zero-shot settings by simply adding the prompt "Let's think step by step," which significantly enhances their performance across a range of tasks. Beyond prompting, CoT has also been explored in model training. For example, Li et al. (2023b) introduced CoT-based knowledge distillation, where the reasoning process is transferred to smaller models, resulting in improved outcomes. Similarly, Ho et al. (2022) fine-tuned smaller models using the reasoning outputs of LLMs. More recently, Liu et al. (2023) proposed a framework that uses LLMs with CoT reasoning to evaluate the quality of generated text within a form-filling paradigm.

3 The FineDialFact Benchmark

Previous works (Gupta et al., 2021; Chen et al., 2024) on dialogue fact verification focus solely on whether the response is factually correct or if there is insufficient information to make a judgment. However, a response may contain factual, incorrect, as well as non-verifiable facts, and only verifying the response is coarse-grained.

To detect the hallucinations in dialogue in a fine-grained way, we aim to verify the atomic fact split by the response. As there are no existing dialogue datasets containing atomic facts, we build one from public dialogue datasets: (1) we generate dialogue response by LLMs as sampling hallucinated exmaples, see Section 3.1 for details; we split the dialogue response into atomic facts based on few-shot learning, as described in Section 3.2; we describe retrieving knowledge, data annotation and evaluation metrics in Sections 3.3, 3.4, and 3.5 respectively.

3.1 Hallucinated Data Sampling

To construct our dataset, two public knowledge-grounded datasets, OpenDialKG and HybriDialogue, are selected. OpenDialKG (Moon et al., 2019) includes a recommendation component focused on movies and books, along with a chit-chat component centred around sports and music. The dataset consists of 1,973 test samples, 9,120 training samples, and 1,962 validation samples. HybriDialogue (Nakamura et al., 2022) is an open-domain dialogue dataset designed for information-seeking conversations, with a train set of 4,359 samples, a validation set of 242 samples, and a test set of 243 samples. However, they only offer dialogue references with factually correct facts, and for fine-grained fact verification, the hallucinated samples are needed. Instead of prompting language models to produce hallucinated content, we guide them to generate responses based on given dialogues to make sure the samples are consistent with the dialogue style.

We adopt various LLMs to generate dialogue responses, including Flan-T5 (Chung et al., 2024)

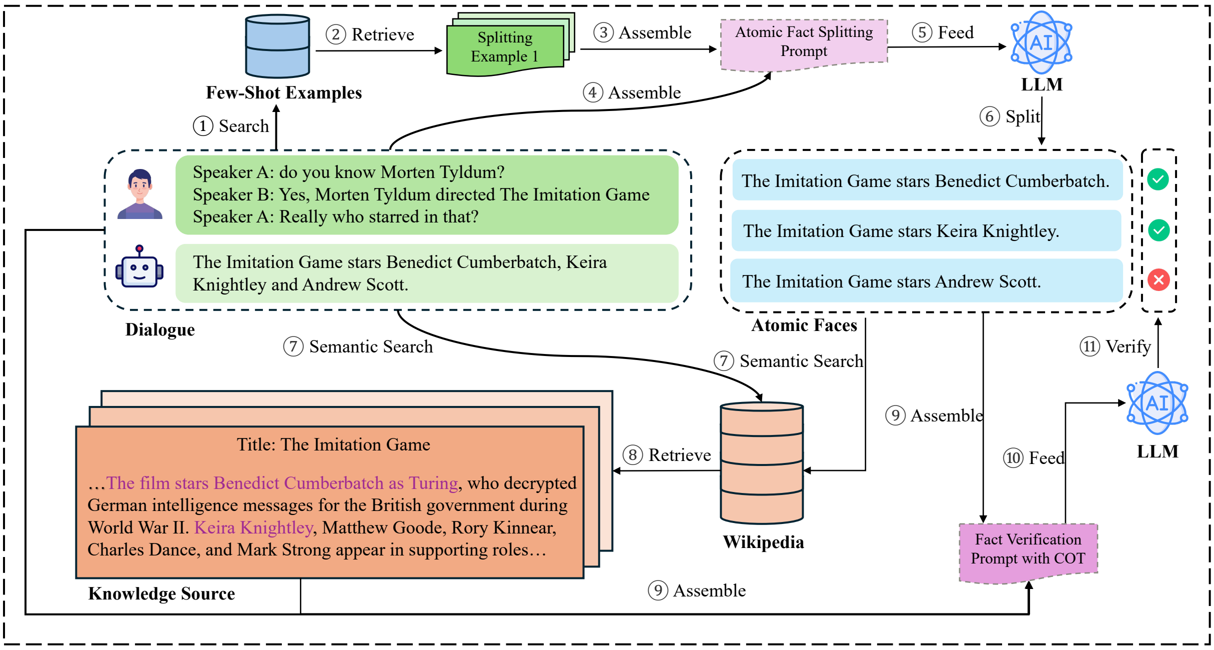

Figure 2: The framework of fine-grained dialogue fact verification. It starts from the dialogue, in which the response is split into several atomic facts. The knowledge source is retrieved from the knowledge database based on the semantic matching of atomic facts and used for precise fact verification.

and Llama3 (Grattafiori et al., 2024), to ensure inclusiveness. The detailed prompt is listed in the Appendix A.

3.2 Atomic Fact Splitting

Min et al. (2023) initially introduced the concept of atomic fact, which represents the smallest unit of a fact. They claimed that the calculation of fact score should be based on atomic facts, instead of a long text.

In this work, we follow their settings to split the response into several pieces of atomic facts by LLMs. To ensure sophisticated splitting, the atomic fact splitting is based on few-shot learning with the retrieval of 2 examples by semantic matching. To pursue replication, the open-source model Llama is employed in our work. The prompt for atomic fact splitting is detailed in Appendix A.

3.3 Knowledge Retriever

Due to the popularity of hallucinations of LLMs, the internal knowledge is unreliable. Therefore, the models rely on external knowledge to verify. We adopt sophisticated Contriever-MS MARCO (Izacard et al., 2021) as our retriever, which is designed by contrastive learning, achieving good performance on document retrieval.

We use Wikipedia as our knowledge source and divide each article into fixed-length passages, as the full article length is often too long for large language models to process effectively.

3.4 Dataset Annotation

We aim to annotate the factual label for each atomic fact by giving a dialogue history and several sources of external knowledge. Additionally, we annotate data from the test sets of both datasets.

After collecting dialogue responses generated by LLMs, we randomly mix them with references from public dialogue datasets. Next, we generate atomic facts from these samples and retrieve the top N corresponding knowledge by the aforementioned Contriever-MS MARCO, based on the semantic matching between Wikipedia texts and the combination of atomic facts and dialogue history.

We recruit several annotators for our task. As these datasets are written in English, only annotators with good English proficiency are considered. Initially, we asked two annotators to independently select the most relevant Wikipedia texts as the knowledge source. Once the Wikipedia texts are selected, the annotators are further asked to verify the atomic fact against the knowledge source and dialogue history, and to assign a factual label: “Supports,” “Refutes,” or “Not Enough Information.” These annotators had no prior experience with AI/LLM and were not aware of hallucination. They were intentionally selected, which helps

Model	Agreement	Kappa	JS
HybriDialogue	0.782	0.615	0.662
OpendialKG	0.788	0.633	0.643

Table 1: Inter-annotator agreement measured by raw agreement and Cohen's Kappa on two datasets. JS is the abbreviation of Jaccard Similarity.

reduce potential bias in the annotations. See Appendix B for more details about annotation instructions.

When the annotation is finished, we assess the similarity of the knowledge source by Jaccard Similarity (Jaccard, 1901). We measure the agreement by Cohen's kappa (Cohen, 1960), which considers chance agreement and is widely used in NLP annotation tasks. If there is a disagreement, we ask a third annotator to choose a factual label from the previous annotators by majority vote.

Table 1 shows the annotation results. The raw agreements of the two datasets are 0.782 and 0.788, respectively, indicating high agreement. Both datasets exhibit Cohen's kappa values above 0.6, indicating a substantial level of agreement. The selection of the knowledge source is reflected in Jaccard Similarity (JS). The JS of both datasets is above 0.6, suggesting a substantial overlap in the selected knowledge content between humans. We finally collected 500 samples from the HybriDialogue dataset and 500 samples from the OpenDialKG dataset, totaling 1,000 samples. The distribution of factual labels is described in Table 2.

Label	HD	OKG	FDF (total)
Supports	181	200	381
Refutes	55	42	97
Not Enough Info	264	258	522
Total	500	500	1,000

Table 2: Distribution for factual labels in FineDialFact (FDF) and by source: HybriDialogue (HD) and OpendialKG (OKG).

3.5 Evaluation Metrics

We use classification metrics to validate the performance of dialogue fact verification, including accuracy, precision, recall and F1-score. Accuracy reflects the overall performance of a classifier, but it may be misleading when dealing with imbalanced data. The F1-score can more realistically reflect performance for imbalanced data.

In addition, we use raw agreement to measure inter-annotator agreement. However, since raw agreement does not account for chance agreement, we also adopt Cohen's Kappa (Cohen, 1960) to evaluate inter-annotator reliability. Furthermore, Cohen's Kappa is employed to assess the model-human agreement between the classifier and human annotators, thereby reflecting agreement beyond chance.

4 Fine-grained Dialogue Fact Verification

We introduce a framework for fine-grained fact verification in dialogue systems, as illustrated in Figure 1. Building on this framework, we propose several CoT baselines to evaluate our dataset. These baselines include zero-shot CoT (Section 4.2), few-shot CoT prompting (Section 4.3), and CoT distillation (Section 4.4).

4.1 Task Definition

We define our task as fine-grained dialogue fact verification. A dialogue is represented as $C = \{c_1, c_2, ..., c_m\}$, where m denotes the number of dialogue turns. The goal is to verify the factual accuracy of the last utterance c_m. This last utterance is decomposed into a set of atomic facts $A = \{a_1, a_2, ..., a_n\}$, where n is the total number of atomic facts. To verify these facts, relevant knowledge is retrieved in the form of passages $T = \{t_1, t_2, ..., t_k\}$, with k indicating the number of retrieved passages. For few-shot learning, we retrieve examples defined as $E = \{e_1, e_2, ..., e_l\}$, where l is the number of examples. Each atomic fact is then classified into one of three labels: "supports", "refutes", and "not enough information", based on the retrieved knowledge.

4.2 Zero-Shot Chain of Thought

Different from traditional fact verification, dialogue history containing a large number of pronoun references should also be considered when verifying facts in dialogue settings, making the task more complex. CoT (Wei et al., 2022) is a kind of method for solving complex tasks. The original CoT requires few reasoning examples. But Kojima et al. (2022) proposed a zero-shot CoT that adding "let's think step by step" into the prompt is able to remarkably improve the LLMs' performance.

CoT has been shown to lead to competitive performance in dialogue fact verification. To verify dialogue facts, we ask the LLM if an atomic fact a_i is factually correct against external knowledge

T and dialogue history $C_{1:m-1}$. And we simply add "think step by step" into the fact verification prompt. The formula is listed as follows:

$$o = \mathcal{M}_d(p_{fact}, a_i, C_{1:m-1}, T)$$

where o denotes output, including factual label and reasoning steps. $\mathcal{M}$ is the LLM for fact verification. p_{fact} is the prompt for verifying facts, described in Appendix A.

4.3 Few-shot CoT Prompting

Few-shot learning is an effective way to improve LLM performance (Brown et al., 2020) without updating weights at inference. Furthermore, Wei et al. (2022) proposed the CoT prompting strategy in a few-shot setting. We follow this setting in our evaluation.

Additionally, we adopt an automated annotation process which enables to annotate 100 samples from the train set which are used as the few shots. Since the annotation process does not contain the CoT process, we adopt GPT-4o to generate the reasoning steps. We retrieve the most relevant samples by semantically matching the atomic facts, defined as follows:

$$o' = \mathcal{M}_d(p_{fact}, a_i, C_{1:m-1}, T, E)$$

where E denotes the retrieved examples and o' is the LLM output based on few-shot learning.

4.4 Reasoning Distillation

Traditional knowledge distillation processes knowledge, usually in the form of labels, from larger to smaller, student models. As we mentioned above, dialogue fact verification is more complex, and relying on teaching labels to smaller models is insufficient.

Unlike the traditional method, we inject reasoning steps when distilling knowledge into student models. Specifically, we request GPT-4o to simulate the human annotation process: select the knowledge source, generate the factual label with the reasoning steps.

After collecting these samples, we fine-tune the smaller models with LoRA (Hu et al., 2021). LoRA is an efficient fine-tuning technique with a few extra parameters and lower computational resources. Another benefit is that it does not change the original LLM weight, fully leveraging the LLMs' strength. We adopt cross-entropy loss to optimize our model, formulated as follows:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(p_i) \tag{1}$$

where N is the number of samples, y denotes the ground truth label and p is the predicted probability for the i-th sample.

5 Experiment

5.1 Baselines

We adopt several LLMs as baselines with various baseline methods to measure the performance of models. The baselines are listed as follows:

- **Flan-T5** (Chung et al., 2022) is a collection of encoder-decoder based LLMs with different model scales. These LLMs are fine-tuned by well-designed instructions, allowing them to perform well in various NLP tasks. In this paper, we adopt the Flan-T5-XXL version.
- **Llama3** (Dubey et al., 2024) is based on a decoder-only architecture, showing strong performance in code generation and multi-turn dialogue generation. We use the Instruct ones in our work.
- **GPT-4o** (Hurst et al., 2024) is an omni-modal, auto-aggressive and closed-source model, achieving state-of-the-art performance in various tasks.

5.2 Experimental Setup

We provide the experimental details about our methodology and datasets. For every atomic fact splitting, we retrieve 3 examples for few-shot learning. The samples used for fine-tuning are from the train set of OpendialKG and generated by GPT-4o, with the number of 3000. We used LoRA to fine-tune our smaller language models, with the settings of rank 32 and alpha 32. We fine-tune the Llama 8B model for 3 epochs with a single 80GB A100 GPU. For the open-source models, the inference with the Llama 70B model requires two 80GB A100 GPUs, and all the other models use one. The experiments were conducted using a fixed random seed of 42, with a single run. All experiments took approximately 120 hours in total.

5.3 Experimental Results and Analysis

We report the dialogue fact verification results on two public datasets, OpendialKG and HybriDialogue, in Table 3 and 4 respectively. We analyze the models' overall performance and their agreement with humans.

Model	Accuracy	Precision	Recall	F1-score	Kappa
			Vanilla		
Flan-T5-XXL	0.606	0.563	0.567	0.560	0.297
Llama-3.1-8B-Instruct	0.512	0.665	0.617	0.498	0.222
Llama-3.3-70B-Instruct	0.822	**0.794**	0.789	0.791	0.687
GPT-4o	**0.836**	0.793	**0.847**	**0.814**	**0.715**
			CoT		
Flan-T5-XXL	0.568	0.540	0.534	0.526	0.218
Llama-3.2-1B-Instruct*	0.812	0.639	0.612	0.604	0.657
Llama-3.2-3B-Instruct*	0.838	0.771	0.706	0.728	0.704
Llama-3.1-8B-Instruct	0.746	0.707	0.580	0.584	0.535
Llama-3.1-8B-Instruct$^\diamond$	0.844	0.796	0.661	0.672	0.715
Llama-3.1-8B-Instruct*	**0.870**	**0.895**	0.762	0.804	**0.761**
Llama-3.3-70B-Instruct	0.836	0.791	0.805	0.797	0.712
GPT-4o	0.864	0.874	**0.823**	**0.843**	0.754
			Few-Shot CoT		
Flan-T5-XXL	0.528	0.440	0.369	0.319	0.064
Llama-3.1-8B-Instruct	0.802	0.756	0.715	0.731	0.647
Llama-3.3-70B-Instruct	0.862	0.857	**0.845**	0.849	0.757
GPT-4o	**0.884**	**0.868**	0.839	**0.851**	**0.792**

Table 3: The test results on the OpendialKG dataset. The bold number means the best performance within the same methods. Kappa means Cohen's Kappa, indicating the inter-agreement between humans and models. Vanilla models refer to those without CoT reasoning. Models marked with $^\diamond$ are fine-tuned exclusively on factual labels, whereas those marked with * are fine-tuned on factual labels augmented with CoT data. This distinction reflects the different types of knowledge involved in the knowledge distillation process.

Overall Performance We first evaluate the performance of fact verification by analyzing accuracy. There is a similar pattern of accuracy between the two datasets: GPT-4o has the highest performance in few-shot CoT methods, whereas the distillation version Llama-3.1-8B-Instruct gets the most significant score among the CoT methods, demonstrating the superiority of CoT distillation.

With CoT, performance of all LLMs improves greatly, except for Flan-T5-XXL, suggesting the latter is unable to reason. GPT-4o and Llama are beneficial with few-shot examples, which indicates they are good few-shot learner. Notably, the student model Llama-3.1-8B-Instruct* is better than the teacher model, GPT-4o, in accuracy on the two datasets. Furthermore, compared to Llama-3.1-8B-Instruct, Llama-3.1-8B-Instruct* achieves a significant improvement of 15% on HybriDialogue and 16.62% on OpendialKG, demonstrating the effectiveness of our distillation method.

Given that accuracy may be an unreliable metric for imbalanced datasets, we also take precision, recall and F1-score into account in our evaluation. There is a general improvement with CoT methods on precision, making fewer false positive errors. After applying few-shot CoT prompting, the Llama models and GPT-4o increase in recall performance, and the Llama 8B Instruct improves substantially, which indicates that this model has the ability to identify more related samples.

When we look at F1-scores, GPT-4o generally achieves the best results with CoT and few-shot CoT on both datasets. Although GPT-4o performs best overall, it can also be observed that the increase is limited with CoT, while other models like Llama have substantial enhancement. Especially for distilled Llama-3.1-8B-Instruct, it rises from 0.498 to 0.804 on the OpendialKG dataset, and from 0.408 to 0.707 on HybriDialoug dataset.

In summary, Flan-T5-XXL demonstrates limited reasoning capabilities, as there is no observed enhancement with CoT prompting. Overall, GPT-4o shows superior performance in vanilla and few-shot CoT learning. Llama 3 models perform better than GPT-4o in CoT. There is a substantial improvement in the distilled model compared to the vanilla one, showing this method's effectiveness.

Model-Human Agreement As the dialogue fact verification lays the foundation to assess the factu-

Model	Accuracy	Precision	Recall	F1-score	Kappa
			Vanilla		
Flan-T5-XXL	0.626	0.561	0.585	0.570	0.374
Llama-3.1-8B-Instruct	0.448	0.577	0.524	0.408	0.164
Llama-3.3-70B-Instruct	**0.730**	**0.684**	0.675	**0.677**	**0.536**
GPT-4o	0.722	0.671	**0.679**	0.673	0.527
			CoT		
Flan-T5-XXL	0.638	0.581	0.587	0.583	0.376
Llama-3.2-1B-Instruct*	0.768	0.841	0.590	0.580	0.572
Llama-3.2-3B-Instruct*	0.776	0.695	0.611	0.624	0.585
Llama-3.1-8B-Instruct	0.706	0.584	0.549	0.530	0.471
Llama-3.1-8B-Instruct$^{\diamond}$	0.790	**0.774**	0.608	0.586	0.621
Llama-3.1-8B-Instruct*	**0.812**	**0.774**	**0.681**	**0.707**	**0.657**
Llama-3.3-70B-Instruct	0.732	0.666	0.659	0.660	0.537
GPT-4o	0.752	0.708	0.673	0.687	0.562
			Few-Shot CoT		
Flan-T5-XXL	0.502	0.376	0.350	0.309	0.016
Llama-3.1-8B-Instruct	0.718	0.714	0.622	0.638	0.504
Llama-3.3-70B-Instruct	0.752	0.727	0.715	0.718	0.574
GPT-4o	**0.800**	**0.767**	**0.734**	**0.748**	**0.647**

Table 4: The test results on the HybriDialogue dataset are shown. Bold numbers indicate the highest performance achieved within each method category.

ality of dialogue response, we report the Cohen's Kappa coefficient, taking into account the agreement by chance, to indicate the agreement between the ground-truth and the model. In this experiment, kappa score above 0.6 means the model has sustainable agreement with humans. Among vanilla models, kappa score of Llama-3.3-70B-Instruct and GPT-4o on OpendialKG dataset exceeds 0.6, showing sustainable agreement with humans. After CoT, GPT-4o and Llama-3.3-70B-Instruct improve to more than 0.6 on two datasets. Llama-3.1-8B-Instruct* performs best with the CoT method. We also see that there is a similar trend with the accuracy on the two datasets: GPT-4o has the best results in few-shot CoT patterns, while Llama-3.1-8B-Instruct* performs best with the CoT method.

In conclusion, HybriDialogue is more challenging as the models' performance has a lower kappa score. CoT methods can improve the kappa score significantly, especially for the distillation CoT.

5.4 Case Study

We present a case study on the Jaywalk example to illustrate the effectiveness of CoT distillation, as detailed in Appendix D. Initially, the Llama-3.1-8B-Instruct model outputs a "Not Enough Information" label. However, after distillation using data generated by GPT-4o, the model is able to produce the correct label. This demonstrates the superiority of the proposed distillation method.

6 Conclusion

We introduce FineDialFact, a novel benchmark dataset for fine-grained fact verification in dialogue. Previous fact verification on dialogue focuses on the response level, which is coarse-grained. To verify dialogue facts in a fine-grained way, we split the response into small pieces of atomic facts, enabling the challenging yet realistic scenario where different facts within a dialogue can have different factual labels. Given that there were no existing related datasets available, we construct the FineDialFact dataset based on coarse-grained dialogue datasets and by generating hallucinating samples, splitting responses into atomic facts, retrieving knowledge, and ultimately recruiting participants for manual annotation.

We also perform benchmarking experiments with CoT baselines. Experimental results show that CoT can greatly improve the models' performance, but also show that the task is far from solved. On the HybriDialogue dataset, the highest F1-score achieved is 0.735, indicating that dialogue fact verification remains a challenging task.

Limitations

While our proposed benchmark makes a significant contribution to fine-grained dialogue fact verification, some limitations remain in our work.

We split the dialogue response into several pieces of atomic facts to verify, gaining more accurate results. However, it increases the cost of using GPUs.

In addition, our current knowledge base relies exclusively on Wikipedia, which presents a limitation, as incorporating additional sources could enhance the robustness of the verification process.

Ethical Statement

Our work involves human annotations; however, the tasks were limited to labelling a predefined range of options, such as selecting factual labels, and did not involve the collection or use of any personal information.

The datasets we used, HybriDialogue and OpendialKG, are publicly available, and no additional personally sensitive information was added in our benchmark.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

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

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70):1–53.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1):37–46.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017):625–630.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Dialfact: A benchmark for fact-checking in dialogue. arXiv preprint arXiv:2110.08222.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. arXiv preprint arXiv:2212.10071.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q{2}: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. arXiv preprint arXiv:2104.08202.

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. arXiv preprint arXiv:2112.09118.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull Soc Vaudoise Sci Nat, 37:547–579.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. arXiv preprint arXiv:2305.11747.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023b. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. arXiv preprint arXiv:2306.14050.

Yanyang Li, Jianqiao Zhao, Michael R Lyu, and Liwei Wang. 2022. Eliciting knowledge from large pre-trained models for unsupervised knowledge-grounded conversation. arXiv preprint arXiv:2211.01587.

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. arXiv preprint arXiv:2303.16634.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251.

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

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In Proceedings of the 57th annual meeting of the association for computational linguistics, pages 845–854.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. Hybridialogue: An information-seeking dialogue dataset grounded on tabular and textual data. arXiv preprint arXiv:2204.13243.

Xuanfan Ni, Hongliang Dai, Zhaochun Ren, and Piji Li. 2023. Multi-source multi-type knowledge exploration and exploitation for dialogue generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 12522–12537.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. arXiv preprint arXiv:2104.07567.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. Finesure: Fine-grained summarization evaluation using llms. arXiv preprint arXiv:2407.00908.

Xingwei Tan, Yi Cai, and Changxi Zhu. 2019. Recognizing conflict opinions in aspect-level sentiment classification with dual attention networks. 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), pages 3426–3431.

David Wan, Koustuv Sinha, Srini Iyer, Asli Celikyilmaz, Mohit Bansal, and Ramakanth Pasunuru. 2024. Acueval: Fine-grained hallucination evaluation and correction for abstractive summarization. In Findings of the Association for Computational Linguistics ACL 2024, pages 10036–10056.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Jifan Yu, Xiaohan Zhang, Yifan Xu, Xuanyu Lei, Xinyu Guan, Jing Zhang, Lei Hou, Juanzi Li, and Jie Tang. 2022. Xdai: A tuning-free framework for exploiting pre-trained language models in knowledge grounded dialogue generation. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 4422–4432.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. arXiv preprint arXiv:2210.07197.

A Prompts

The prompts for dialogue response generation, atomic facts splitting and dialogue fact verification are listed in these tables 5, 6, and 7.

Prompt for Dialogue Response Generation
Dialogue: {Dialogue History} **Instruction:** Given the above dialogue, please respond to the input below and ensure the response is fluent and fact-consistent in English. **Input:** {The utterance of Speaker A} **Response:**

Table 5: The prompt for dialogue response generation.

Prompt for Atomic Fact Splitting
Examples: {Retrieved Examples} If the following input is an incomplete sentence or a phrase, please output it exactly as it is. Otherwise, if it is a complete sentence, split it into atomic sentences based only on the given information, without adding any additional information or making inferences. **Input:** {Response} **Output:** {Atomic facts}

Table 6: The prompt for atomic fact splitting.

B Annotation Instruction

The details of the annotation instruction are listed in Table 8. Before annotation, we have fully informed the participants that the annotated data will be used in our research and obtained their consent.

C Dataset License

The OpendialKG (Moon et al., 2019) dataset is licensed under the CC BY-NC 4.0 license, which permits research use. Similarly, the HybriDialogue (Nakamura et al., 2022) dataset is available under the MIT license, which also supports use in research contexts.

D Case Study

The case study example regarding CoT distillation is detailed in Table 9.

Prompt for Dialogue Fact Verification

{Demonstrations}

Instruction:

The statement is part of a response in a dialogue. Evaluate the statement strictly based on the provided knowledge source and dialogue history only.
If the statement is not a factual claim (e.g., opinion, question, or unclear assertion), output: `"not enough information."`
If it is a factual claim:

- Output **true** if the statement is directly supported by evidence in the knowledge source or dialogue history.
- Output **false** if the statement is directly contradicted by the knowledge source or dialogue history.
- Output **not enough information** if there is no direct evidence for or against the statement.

Important:
Do not use your internal knowledge or make inferences.
Please think step by step and output your final answer.
Evidence: {Knowledge Source}
Dialogue History: {Dialogue History}
Statement: {Atomic Fact}
Output:

Table 7: The prompt for our dialogue fact verification. The prompt can be used for vanilla, CoT and few-shot CoT by adjusting the prompt slightly.

Human Annotation Instructions

The task aims at annotating dialogue factual responses. For each sample, we provide you with a dialogue, several pieces of evidence, and two labels—factual claim and factual label. Your task is to select the most relevant pieces of evidence (as much as possible) and determine the labels.
There is a list of samples containing dialogue and evidence. Our goal is to select evidence for the last utterance and identify if the last utterance is verifiable or non-verifiable. You need to use the annotation tool to:
1. Factual Claim Discrimination
First, you have to determine whether the last utterance is a factual claim. A factual claim normally contains:

- Specific, verifiable information that can be proven true or false
- Statements about events, measurements, statistics, or observable phenomena
- References to dates, times, people, places, or quantities
- Content that could be checked against reliable sources or evidence
- Statements that are objective rather than expressing opinions or preferences

If it is a factual claim, select **[Verifiable]** and proceed to step 2. Otherwise, select **[Non-Verifiable]** and assign the factual label as **[Not Enough Information]**.
2. Evidence Selection
Manually select evidence for the last utterance from Speaker B.
3. Claim Verification

- If the utterance is an independent atomic fact, verify it using the selected evidence directly.
- If it involves coreference to earlier dialogue, use both the selected evidence and previous dialogue to verify it.

Finally, assign the **Factual Label**:

- **Supports**: The evidence supports the factual claim.
- **Refutes**: The evidence contradicts the factual claim.
- **Not Enough Information**: Evidence is missing or insufficient.

Note: If the response is irrelevant to the context, treat it as a standalone factual claim.
Summary of Options:
1. Factual Claim
NON-VERIFIABLE: No verifiable factual info; includes personal opinions or private info.
VERIFIABLE: Contains verifiable factual info checkable via background corpus (e.g., Wikipedia).
2. Factual Label
Supports: Evidence supports the factual claim.
Refutes: Factual claim contradicts the evidence.
Not Enough Info: No or insufficient evidence to verify the claim.

Table 8: The instructions for dialogue factual annotation.

Case Study 1	
Evidence	Jaywalk (foaled March 30, 2016 in Kentucky) is a Thoroughbred filly racehorse who in 2018 won the $2,000,000 Breeders' Cup Juvenile Fillies, the world's richest race of its class. She is trained by John Servis for owners Cash is King, LLC (Charles J. Zacney, et al.) and D J Stable, LL (Leonard C. Green).
Speaker A	Who created the Eclipse award
Speaker B	The Daily Racing Form, the Thoroughbred Racing Associations, and the National Turf Writers Association all joined forces in 1971 to create the Eclipse Award
Speaker A	Who is the Owner of Jaywalk
Speaker B(Atomic Fact)	Jaywalk won the 2021 American Champion Two-Year-Old Filly award.
Llama-3.1-8B-Instruct	**Decision :Not Enough Information** reasoning: there is no information about jaywalk winning the 2021 american champion two-year-old filly award in the given evidence or dialogue history.
Llama-3.1-8B-Instruct*	**Decision: Refutes** **chain of thought**: 1. the claim states that jaywalk won the 2021 american champion two-year-old filly award. 2. the evidence provided describes jaywalk as a filly who won the breeders' cup juvenile fillies in 2018, which is a prestigious race. 3. however, the evidence does not mention anything about jaywalk winning the 2021 american champion two-year-old filly award. 4. the evidence only confirms her achievement in 2018, not in 2021. 5. therefore, the evidence does not support the claim about the 2021 award. **output**: false
Ground truth	**Refutes**

Table 9: An example of how CoT distillation works in dialogue fact verification.