FineDialFact: A Benchmark for Fine-Grained Dialogue Fact Verification

Xiangyan Chen[1], **Yufeng Li**[1], **Yujian Gan**[2], **Arkaitz Zubiaga**[1], **Matthew Purver**[1, 3]
[1]Queen Mary University of London, UK
[2] Queen's University Belfast, UK
[3] Institut Jožef Stefan, Slovenia
{xiangyan.chen, yufeng.li, a.zubiaga, m.purver}@qmul.ac.uk, y.gan@qub.ac.uk

Abstract

Large language models are known to produce hallucinations — factually incorrect or fabricated information — which poses significant challenges for many natural language processing 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 non-verifiable facts, making the use of a single 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 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 0.74, indicating that the benchmark remains a challenging task for future research. We release our dataset and code at https://github.com/XiangyanChen/FineDialFact.

Keywords: large language models, fine-grained dialogue fact verification, dialogue hallucination detection

1. Introduction

In recent years, large language models (LLMs) have demonstrated impressive capabilities across a wide range of tasks (Zhao et al., 2023). 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 (Ji et al., 2023).

Previous approaches to hallucination detection in dialogue systems mainly rely on human evaluation (Ni et al., 2023; Li et al., 2022; Shuster et al., 2021; Yu et al., 2022), which is time-consuming and labour-intensive. Recently, automatic methods have been proposed, including uncertainty estimation (Farquhar et al., 2024) and fact verification (Chen et al., 2024). However, uncertainty estimation often fails when the model is overconfident in hallucinated content. Alternatively, our work focuses on direct fact verification. Existing fact verification methods for dialogue systems (Chen et al., 2024; Gupta et al., 2021) verify responses using external knowledge and dialogue context, and output one of three labels: *Supports*, *Refutes*, or *Not Enough Information*. Yet, these methods operate only at the response level, ignoring that a single response may contain all factual, hallucinated, and non-verifiable information. As shown in Figure 1, labelling the entire response as incorrect is overly coarse since it also contains accurate facts.

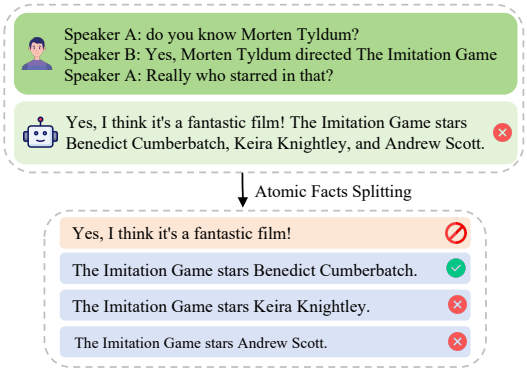

Figure 1: An example of the response-level based dialogue fact verification and fine-grained one. The difference between them is that the latter is based on the atomic facts. The prohibition symbol means it is not a verifiable factual claim.

To address the above limitation, we systematically study fine-grained fact verification for dialogue systems and offer a benchmark, named *FineDialFact*. We process the dialogue response into small, verifiable sentences, called atomic facts. Each atomic fact expresses exactly one proposition that can be judged and then is verified independently using external knowledge and LLMs. As no existing datasets are available, we construct a dataset by extending two public dialogue datasets, OpenDialKG and HybriDialogue, and report their inter-annotator agreement using Cohen's Kappa (Cohen, 1960). To evaluate the dataset, we provide a set of metrics: accuracy, precision, recall, F1-score, Geometric Mean (G-mean), and Cohen's Kappa, which

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2798–2811
11–16 May 2026.

measure performance from different perspectives.

In addition, we evaluate a series of Chain-of-Thought (CoT) based approaches using different LLMs, such as Llama3 (Dubey et al., 2024), Deepseek-R1 (Guo et al., 2025), QwQ (Team, 2025), and GPT (Hurst et al., 2024). These CoT approaches include zero-shot CoT, few-shot CoT, and CoT distillation. For zero-shot CoT, we simply add a reasoning prompt. 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, fact verification on the manually annotated HybriDialogue dataset remains challenging, with the best F1-score of 0.74 achieved by Llama-3.3-70B.

The contributions can be listed as follows:

1. We delve into fine-grained fact verification for dialogue systems by developing a novel benchmark named FineDialFact. To the best of our knowledge, this is the first systematic research on the fine-grained factuality evaluation of dialogue systems.
2. We provide a newly constructed dataset, including manually and automatically annotated data, to evaluate 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 human-annotated HybriDialogue dataset is more challenging, and the highest score, achieved by Llama-3.3-70B, is only 0.74, 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 labour-intensive, highlighting the need for effective automatic evaluation methods. Chen et al. (2024) and Gupta et al. (2021) addressed a critical gap with a dialogue-level fact-verification and hallucination detection benchmark that extends beyond factuality.

Despite these advances, current methods still struggle with dialogue responses that mix correct and incorrect information.

2.2. Fine-grained Detection

Inspired by the challenges observed in dialogue hallucination detection, we further review fine-grained detection techniques developed in general domains. Aspect-based sentiment analysis acknowledges the possibility of having both positive and negative sentiments in the same sentence (Tan et al., 2019). Song et al. (2024) and Wan et al. (2024) introduced fine-grained techniques for detecting hallucinations in text summarisation, allowing for more accurate identification of factual errors. 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. Similarly, Mitra et al. (2024) provided a fact verification benchmark aimed at splitting claims into sub-claims and analysed the importance of the quality of sub-claims.

Nevertheless, the above studies have not addressed the unique challenges of the dialogue domain, where fact verification must account for the evolving conversational context, making the task more complex. To this end, we introduce a fine-grained approach for dialogue fact verification.

2.3. Chain of Thought

Since the rise of LLMs, there has been growing interest in applying them to NLP tasks. The CoT approach (Wei et al., 2022) improves performance on complex tasks by introducing intermediate reasoning steps. To reduce the need for hand-crafted few-shot examples, Zhang et al. (2022) proposed automatically collecting examples via clustering. Kojima et al. (2022) showed that zero-shot CoT can also work well by prompting with "Let's think step by step." CoT has also been applied in training, such as in CoT-based knowledge distillation (Li et al., 2023), where transferring reasoning to smaller models boosts performance.

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 has insufficient information to make a judgment. However, a response may contain factually correct and incorrect facts, as well as non-verifiable factual claims, and only verifying the response is coarse-grained.

To detect hallucinations in dialogue at a fine-grained level, we aim to verify each verifiable atomic fact extracted from the response. As there are no existing dialogue datasets containing atomic facts, we build one extending public dialogue datasets: (1) we generate dialogue response by LLMs as sampling hallucinated examples, see Section 3.1 for details; (2) we split the dialogue response into

atomic facts based on few-shot learning, as described in Section 3.2; (3) we describe retrieving knowledge (Section 3.3), manual (Section 3.4) and automated data annotation (Section 3.5), and evaluation metrics (Section 3.6).

3.1. Hallucinated Data Sampling

To construct our dataset, we select two public knowledge-grounded datasets: OpenDialKG and HybriDialogue. 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. HybriDialogue (Nakamura et al., 2022) is an open-domain dialogue dataset designed for information-seeking conversations. However, it only offers dialogue references with factually correct facts, and for fine-grained fact verification, the hallucinated samples are needed. Instead of prompting LLMs to produce hallucinated content, we guide them to generate responses based on given dialogues, ensuring the samples are consistent with the dialogue style.

We adopt various LLMs to generate dialogue responses, such as Llama-3.1-8B-Instruct, Flan-T5-XXL (Chung et al., 2024), to ensure inclusiveness.

3.2. Atomic Fact Splitting

We define atomic facts as minimal, independent propositions in line with Min et al. (2023), and we verify each unit separately with external evidence and an LLM-based judge.

In this work, we follow Min et al. (2023) setup to decompose responses into multiple atomic facts using LLMs. The splitting process relies on few-shot learning, with two examples retrieved through BM25 (Robertson and Zaragoza, 2009). The atomic splitting is construction-agnostic and is guided by the few-shot demonstrations. The dialogue scenario contains some non-verifiable atomic claims, such as opinions, which are not considered to be verified further. For reproducibility, the open-source model Llama-3-70B-Instruct is used in the atomic fact splitting.

We randomly selected 100 samples from the HybriDialogue dataset to assess the quality of atomic fact splitting. Atomic facts were rated as Good (1), Acceptable (0.5), or Bad (0) based on accuracy, completeness, and clarity. The raw agreement between annotators is 0.85, with a Cohen's kappa of 0.562, indicating moderate agreement. The average score is 0.883, showing high-quality atomic fact splitting.

3.3. Knowledge Retriever

Due to the ubiquity of LLM hallucinations, 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, dividing each article into fixed-length passages, since full articles are often too long for LLMs to process.

Task	Humans				GPT-4o			
	HD		ODKG		HD		ODKG	
	Agr.	Kappa	Agr.	Kappa	Agr.	Kappa	Agr.	Kappa
Factual Claim	0.865	0.661	0.865	0.692	0.864	0.657	0.848	0.687
Factual Label	0.722	0.546	0.735	0.533	0.754	0.594	0.840	0.717

Table 1: Comparison of agreement for two annotation tasks: The left panel shows human inter-annotator agreement for HybriDialogue (HD) and OpenDialKG (ODKG), and the right panel shows GPT-4o with gold labels. Metrics include raw agreement (Agr.) and Cohen's kappa.

Task	Humans		GPT-4o	
	HD	ODKG	HD	ODKG
Evidence Selection	0.577	0.556	0.642	0.590

Table 2: Annotation overlap (Jaccard Similarity) results for the evidence selection task. The left shows inter-annotator overlap, and the right shows overlap between ground truth and GPT-4o.

3.4. Manual Data Annotation

After collecting dialogue responses generated by LLMs, we randomly mix them with reference responses from public dialogue datasets. We then extract atomic facts from these samples and retrieve the top N relevant knowledge passages from Wikipedia using the Contriever-MS MARCO retriever. Retrieval is performed via semantic matching between Wikipedia passage embeddings and a query embedding, using cosine similarity as the similarity metric. The query is constructed from the atomic facts and the dialogue history.

We then ask two annotators to follow a three-step annotation process: first, they assess the verifiability of each factual claim; second, they select the most relevant Wikipedia passages as evidence for each verifiable atomic fact; third, they verify each atomic fact against the selected evidence and dialogue history, assigning one of three labels — *Supports*, *Refutes*, or *Not Enough Information*. All annotators perform these steps independently. Since the datasets are in English, only annotators with good English proficiency are selected. None of them have prior experience with AI or LLMs, nor familiarity with hallucination phenomena, which helps

Label	Humans			GPT-4o		
	HD	ODKG	FDF	HD	ODKG	FDF
Supports	181	200	381	1,833	2,153	3,986
Refutes	55	42	97	249	201	450
NEI	134	83	217	1,896	1,723	3,619
Total	370	325	695	3,978	4,077	8,055

Table 3: Comparison of FineDialFact (FDF) label distributions in human and automated datasets. NEI stands for Not Enough Information. HD and ODKG refer to the datasets HyriDialogue and OpenDialKG, respectively.

reduce potential biases in the annotations.

After annotation, we assessed the similarity of the knowledge source by Jaccard Similarity (JS) (Jaccard, 1901). We measured the agreement using Cohen's Kappa (Cohen, 1960), which considers chance agreement and is widely used in NLP annotation tasks. When there was a disagreement, we asked a third annotator to choose a factual label by majority vote among the previous annotators.

We sampled 500 atomic claims each from HybriDialogue and OpenDialKG (1,000 total). As shown in Table 1, factual-claim annotation attains an agreement of 0.865 on both datasets with substantial consistency (Cohen's kappa > 0.6). Fact labelling yields agreements of 0.722 and 0.735 with moderate consistency (kappa > 0.5). Human knowledge-source choices have JS of about 0.6 on both datasets, indicating moderate overlap. After filtering non-factual claims, the distribution of factual labels is reported in Table 3.

3.5. Automated Data Annotation

As an efficient approach to increase dataset size, we adopt GPT-4o for automated data annotation following a three-step process: detecting verifiable claims, evidence selection, and fact verification.

For identifying verifiable factual claims and selecting evidence, we use zero-shot prompting, while few-shot learning is applied for fact verification, as detailed in Section 4.3. We first assess the LLM's performance on these three tasks (see Table 1 and 2); the results show that GPT-4o achieves a Cohen's kappa and JS of approximately 0.6, indicating substantial agreement with the ground truth and demonstrating its reliability.

We automatically annotated the atomic facts in 500 dialogues from each dataset. Table 3 presents the distribution of the atomic facts. We use GPT-4o to identify the verifiable factual claims and select evidence per dialogue response. These annotations allow us to analyse the relationship between response-level and atomic-level fact verification.

3.6. 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 and Geometric Mean (G-Mean) can more realistically reflect performance for imbalanced data. The G-Mean is calculated based on the recall scores of different classes. In addition, Cohen's kappa is employed to assess the model-human agreement between the classifier and annotator, 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 CoT baselines to evaluate the datasets, including 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 $\mathcal{U} = \{\mathrm{U}_1, \mathrm{U}_2, ..., \mathrm{U}_m\}$, where m denotes the number of dialogue turns. The goal is to verify the factual accuracy of the last utterance U_m. This last utterance is decomposed into a set of atomic facts $\mathcal{A} = \{a_1, a_2, ..., a_n\}$, where n is the total number of atomic facts. To verify these facts, relevant knowledge is retrieved in the form of passages $\mathcal{T} = \{t_1, t_2, ..., t_k\}$, with k indicating the number of retrieved passages. For few-shot learning, we retrieve examples defined as $\mathcal{E} = \{e_1, e_2, ..., e_l\}$, 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 prompting strategy for solving complex tasks. The original CoT requires few reasoning examples. But Kojima et al. (2022) proposed a zero-shot CoT, showing that adding "let's think step by step" to the prompt can remarkably improve LLM 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

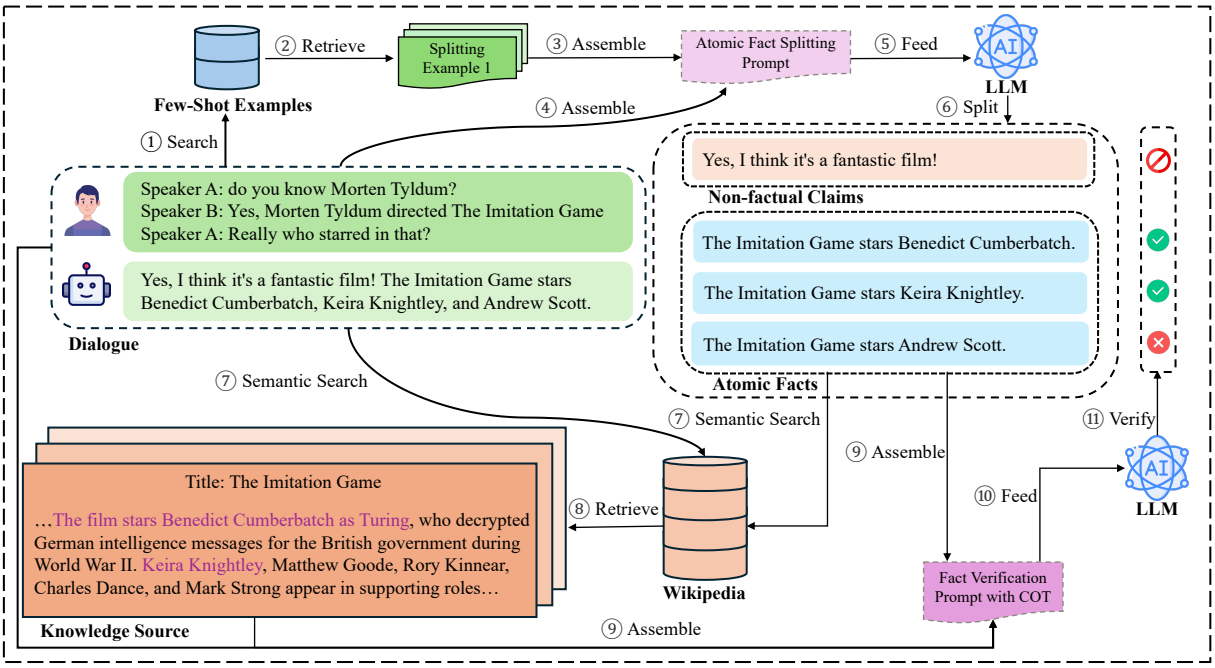

Figure 2: The framework for fine-grained dialogue fact verification. Starting from a dialogue, an LLM splits the response into several atomic facts using few-shot prompting. For each atomic fact, relevant evidence passages are retrieved from Wikipedia via semantic search. An LLM then verifies each atomic fact against the retrieved evidence and outputs the factual label.

is factually correct against external knowledge $\mathcal{T}$ and dialogue history $\mathcal{U}_{1:m-1}$. And we simply add "think step by step" into the fact verification prompt. The formula is listed as follows:

$$(o_i^{\text{reason}}, o_i^{\text{label}}) = \mathcal{M}_\theta(p_{\text{fact}}, a_i, \mathcal{U}_{1:m-1}, \mathcal{T}), \quad (1)$$

where o_i^{reason} and o_i^{label} denote outputs of the reasoning process and factual label, including factual label and reasoning steps. $\mathcal{M}$ is the LLM for fact verification. p_{fact} is the prompt template for verifying facts,

4.3. Few-shot Chain-of-Thought 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, which we follow in our evaluation.

Additionally, we employ an automated annotation process, which enables us to annotate 100 samples from the training set to construct an exemplar pool. Since the annotation process does not contain the CoT process, we adopt GPT-4o to generate the reasoning steps.

Given an atomic fact a_i, we retrieve the most relevant few-shot exemplars $\mathcal{E}$ by semantic matching between a_i and the exemplar pool. We then perform few-shot CoT fact verification conditioned on the dialogue history $\mathcal{U}_{1:m-1}$, the external knowledge $\mathcal{T}$, and the retrieved exemplars $\mathcal{E}$, producing both a rationale and a veracity label:

$$(o_i^{\text{fs,reason}}, o_i^{\text{fs,label}}) = \mathcal{M}_\theta(p_{\text{fact}}, a_i, \mathcal{U}_{1:m-1}, \mathcal{T}, \mathcal{E}), \quad (2)$$

where $o_i^{\text{fs,reason}}$, $o_i^{\text{fs,label}}$ are few-shot LLM outputs. We use the same prompt template p_{fact} described in Section 4.2 but add the exemplars for few-shot prompting.

4.4. Reasoning Distillation

Traditional knowledge distillation processes knowledge, usually in the form of labels, from larger to smaller 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 collect training samples by requesting GPT-4o to simulate the automated annotation process: identity verifiable factual claims, select the knowledge source, and generate the factual label with the reasoning steps.

After collecting these samples, We fine-tune the smaller models with LoRA (Hu et al., 2021), which adapts the original weights via low-rank matrices. We optimise the student model using a combination of label-level and reasoning-level losses, defined

Model	HybriDialogue						OpenDialKG					
	Acc.	Prec.	Rec.	F1	Kappa	G-mean	Acc.	Prec.	Rec.	F1	Kappa	G-mean
Vanilla												
Mistral-7B-Instruct-v0.3	0.646	0.641	0.573	0.572	0.379	0.513	0.772	0.723	0.660	0.680	0.543	0.626
Llama-3.1-8B-Instruct	0.557	0.493	0.496	0.429	0.208	0.242	0.714	0.672	0.594	0.542	0.377	0.300
Llama-3.3-70B-Instruct	0.741	0.725	**0.684**	0.698	0.568	**0.669**	0.822	0.808	0.776	0.784	0.675	0.771
Gemma-3-27B-it	**0.757**	**0.750**	0.682	**0.699**	**0.591**	0.652	0.831	0.799	0.774	0.780	0.691	0.768
Qwen3-32B	0.730	0.697	0.660	0.671	0.541	0.637	0.849	**0.856**	0.774	0.806	0.701	0.761
QwQ-32B	0.714	0.666	0.666	0.659	0.527	0.652	**0.852**	0.812	**0.802**	**0.806**	**0.722**	**0.796**
GPT-4o	0.711	0.666	0.665	0.665	0.524	0.656	0.788	0.745	0.783	0.760	0.627	0.782
CoT												
Mistral-7B-Instruct-v0.3	0.659	0.648	0.576	0.585	0.404	0.529	0.769	0.721	0.650	0.675	0.537	0.619
DeepSeek-R1	0.727	0.727	0.685	0.694	0.552	0.672	0.778	0.766	0.748	0.735	0.618	0.735
Llama-3.1-8B-Instruct	0.651	0.596	0.515	0.493	0.389	0.311	0.668	0.574	0.489	0.474	0.355	0.322
Llama-3.1-8B-Instruct$^{\diamond}$	0.708	0.722	0.570	0.540	0.495	0.334	0.760	0.725	0.606	0.625	0.520	0.526
Llama-3.1-8B-Instruct*	0.759	**0.759**	0.666	0.671	0.597	0.594	0.782	0.759	0.700	0.694	0.611	0.655
Llama-3.3-70B-Instruct	0.743	0.735	0.683	0.699	0.571	0.665	0.840	**0.816**	0.795	0.798	0.710	0.789
Gemma-3-27B-it	**0.770**	0.748	0.700	0.716	**0.612**	0.678	0.843	0.799	0.776	0.785	0.708	0.768
Qwen3-32B	0.759	0.719	0.687	0.696	0.599	0.660	0.849	0.812	0.819	0.812	0.728	0.817
QwQ-32B	0.759	0.716	**0.725**	**0.720**	0.607	**0.720**	**0.855**	0.807	**0.844**	**0.824**	**0.740**	**0.843**
GPT-4o	0.746	0.739	0.699	0.708	0.582	0.684	0.797	0.793	0.800	0.781	0.651	0.797
Few-Shot CoT												
Mistral-7B-Instruct-v0.3	0.643	0.677	0.577	0.574	0.372	0.514	0.766	0.740	0.671	0.672	0.529	0.619
DeepSeek-R1	0.719	0.732	0.669	0.680	0.538	0.647	0.769	0.787	0.746	0.735	0.605	0.732
Llama-3.1-8B-Instruct	0.708	0.703	0.617	0.633	0.501	0.568	0.791	0.736	0.687	0.700	0.606	0.657
Llama-3.3-70B-Instruct	**0.776**	**0.752**	**0.730**	**0.740**	**0.627**	**0.722**	**0.862**	**0.833**	**0.828**	**0.830**	0.744	0.825
Gemma-3-27B-it	**0.776**	0.747	0.717	0.729	0.623	0.703	0.858	0.828	0.783	0.803	0.727	0.774
Qwen3-32B	0.724	0.711	0.663	0.672	0.544	0.636	0.825	0.799	0.805	0.796	0.688	0.804
QwQ-32B	0.768	0.721	0.716	0.718	0.616	0.705	0.858	0.793	0.825	0.807	**0.745**	0.824
GPT-4o	0.754	0.737	0.704	0.712	0.594	0.688	0.840	0.824	0.827	0.817	0.717	**0.826**

Table 4: Comparison of model performances on HybriDialogue and OpenDialKG datasets under Vanilla, CoT, and Few-Shot CoT settings. Kappa means Cohen's kappa, indicating the inter-agreement between humans and models. Vanilla models refer to those without CoT reasoning. Llama-3.1-8B-Instruct$^{\diamond}$ is fine-tuned exclusively on factual labels, whereas Llama-3.1-8B-Instruct* is fine-tuned on factual labels augmented with reasoning steps. The best results are bolded in each category.

as:

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \mathcal{L}_{\text{reason}}. \quad (3)$$

The label loss is the standard cross-entropy between the teacher-generated labels $y^{teacher}$ and the student predictions $p^{student}$:

$$\mathcal{L}_{\text{label}} = -\sum_{i=1}^{N} y_i^{\text{teacher}} \log p_i^{\text{student}}, \quad (4)$$

and the reasoning loss distils the reasoning sequences from the teacher to the student. We implement it as a token-level cross-entropy over the reasoning text generated by the teacher:

$$\mathcal{L}_{\text{reason}} = -\sum_{i=1}^{N} \sum_{j=1}^{L_i} \log P_\theta^{\text{student}}(w_{i,j} \mid w_{i,<j}), \quad (5)$$

where N is the number of samples, L_i is the sequence length, y is the ground-truth label, and p is the predicted probability for the i-th sample. $w_{i,j}$ denotes the j-th token in the i-th reasoning sequence.

5. Experiment

5.1. Baselines

We adopt several LLMs as baselines with various baseline methods to measure the performance of models. The LLMs include Mistral-7B (Albert et al., 2023), Llama3 (Dubey et al., 2024), Deepseek-R1 (Guo et al., 2025), Qwen3 (Yang et al., 2025), QwQ (Team, 2025), Gemma3 (Team et al., 2025), Gemini (Team et al., 2024), GPT (Hurst et al., 2024).

5.2. Experimental Setup

The samples used for fine-tuning are from the train set of OpenDialKG and generated by GPT-4o, with a total of 1429. We used LoRA to fine-tune our smaller language models, with the settings of rank 32 and alpha 32. We fine-tune the Llama 3 8B models for 3 epochs with a single 80GB A100 GPU. For the open-source models, the inference with the Llama 3 70B models 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.

Model	HybriDialogue						OpenDialKG					
	Acc.	Prec.	Rec.	F1	Kappa	G-mean	Acc.	Prec.	Rec.	F1	Kappa	G-mean
Vanilla												
Mistral-7B-Instruct-v0.3	0.631	0.666	0.654	0.585	0.355	0.581	0.677	0.660	0.669	0.580	0.393	0.597
Llama-3.1-8B-Instruct	0.532	0.631	0.614	0.451	0.200	0.342	0.577	0.633	0.598	0.420	0.205	0.303
Llama-3.3-70B-Instruct	**0.855**	0.816	**0.851**	**0.827**	**0.743**	**0.847**	**0.811**	**0.741**	**0.804**	**0.744**	**0.650**	**0.791**
Gemma-3-27B-it	0.845	**0.818**	0.846	0.822	0.726	0.840	**0.811**	0.737	0.789	0.737	0.649	0.776
QwQ-32B	0.751	0.700	0.786	0.684	0.578	0.758	0.710	0.665	0.718	0.600	0.474	0.660
Qwen3-32B	0.758	0.763	0.777	0.737	0.573	0.752	0.753	0.733	0.736	0.689	0.532	0.703
CoT												
Mistral-7B-Instruct-v0.3	0.654	0.663	0.649	0.602	0.390	0.597	0.698	0.667	0.679	0.608	0.431	0.625
Deepseek-R1	0.871	0.828	0.858	0.842	0.769	0.858	**0.856**	**0.782**	0.828	**0.801**	0.736	0.826
Llama-3.1-8B-Instruct	0.740	0.690	0.579	0.594	0.519	0.471	0.734	0.637	0.543	0.556	0.479	0.419
Llama-3.1-8B-Instruct$^{\diamond}$	0.727	0.793	0.545	0.553	0.490	0.378	0.766	0.735	0.641	0.666	0.548	0.600
Llama-3.1-8B-Instruct*	0.831	0.820	0.776	0.795	0.694	0.769	0.817	0.765	0.755	0.759	0.658	0.748
Llama-3.3-70B-Instruct	**0.872**	0.828	**0.874**	**0.843**	**0.773**	**0.871**	0.839	0.776	0.832	0.785	**0.702**	0.824
Gemma-3-27B-it	0.854	**0.831**	0.870	0.836	0.741	0.862	0.824	0.769	0.802	0.764	0.671	0.789
QwQ-32B	0.843	0.768	0.872	0.789	0.729	0.865	0.828	0.750	**0.846**	0.760	0.687	**0.833**
Qwen3-32B	0.842	0.793	0.847	0.811	0.721	0.843	0.816	0.757	0.815	0.770	0.660	0.807
Few-Shot CoT												
Mistral-7B-Instruct-v0.3	0.590	0.625	0.640	0.531	0.296	0.537	0.646	0.637	0.634	0.539	0.332	0.542
Deepseek-R1	0.873	0.832	0.859	**0.845**	**0.774**	0.859	**0.857**	**0.802**	0.817	**0.809**	**0.736**	0.814
Llama-3.1-8B-Instruct	0.816	0.806	0.790	0.788	0.671	0.781	0.801	0.740	0.750	0.730	0.626	0.736
Llama-3.3-70B-Instruct	0.853	**0.812**	0.862	0.824	0.741	0.856	0.831	0.776	0.827	0.782	0.685	0.816
Gemma-3-27B-it	0.828	0.804	0.843	0.802	0.697	0.831	0.805	0.774	0.797	0.757	0.633	0.778
QwQ-32B	**0.856**	0.795	**0.878**	0.817	0.749	**0.872**	0.828	0.759	**0.833**	0.766	0.683	**0.822**
Qwen3-32B	0.853	0.805	0.843	0.821	0.738	0.842	0.827	0.767	0.817	0.782	0.680	0.812

Table 5: Comparison of model performance on the GPT-4o–annotated HybriDialogue and OpenDialKG datasets. Llama-3.1-8B-Instruct$^{\diamond}$ is fine-tuned solely on factual labels, while Llama-3.1-8B-Instruct* is fine-tuned on factual labels enriched with reasoning steps. We bold the best result in each category.

Mode	Model	Response-level			Atomic-Fact-level		
		SUPPORTS	REFUTES	NEI	SUPPORTS	REFUTES	NEI
OpenDialKG Dataset							
Vanilla	Mistral-7B-Instruct-v0.3	441 (0.641)	142 (0.206)	105 (0.153)	3137 (0.769)	394 (0.097)	546 (0.134)
Few-Shot CoT	Qwen3-32B	271 (0.394)	145 (0.211)	272 (0.395)	2423 (0.594)	277 (0.068)	1377 (0.338)
Few-Shot CoT	Llama-3.3-70B-Instruct	288 (0.419)	159 (0.231)	241 (0.350)	2578 (0.632)	297 (0.073)	1202 (0.295)
HybriDialogue Dataset							
Vanilla	Mistral-7B-Instruct-v0.3	658 (0.629)	173 (0.165)	215 (0.206)	3027 (0.761)	350 (0.088)	601 (0.151)
Few-Shot CoT	Qwen3-32B	335 (0.320)	166 (0.159)	545 (0.521)	1944 (0.489)	299 (0.075)	1735 (0.436)
Few-Shot CoT	Llama-3.3-70B-Instruct	399 (0.381)	183 (0.175)	464 (0.444)	2205 (0.554)	330 (0.083)	1443 (0.363)

Table 6: Number and proportions of *Supports*, *Refutes*, and *NEI (Not Enough Information)* predicted at two levels by several LLMs on the GPT-4o-annotated HybriDialogue and OpenDialKG datasets.

5.3. Results on Human-Annotated Datasets

We report dialogue fact verification results on two manually annotated datasets, OpenDialKG and HybriDialogue (Table 4).

In the vanilla setting, Llama-3.3-70B-Instruct and Gemma-3-27B-it deliver the best overall results across both datasets. On HybriDialogue, Gemma-3-27B achieves the highest accuracy and Cohen's kappa, while on OpenDialKG, QwQ-32B achieves the best F1 score, indicating stronger generalisation on that dataset.

With CoT prompting, all models improve across metrics on both datasets, particularly recall and Cohen's kappa. Under CoT, Gemma-3-27B-it and QwQ-32B are the top performers; QwQ-32B yields the highest F1 (0.720 on HybriDialogue and 0.740 on OpenDialKG), demonstrating that CoT benefits LLM understanding. Generally, a higher G-mean with CoT indicates a more balanced per-class recall under class imbalance, suggesting improvements are not driven solely by the majority class.

Regarding distillation, Llama-3.1-8B-Instruct$^{\diamond}$ denotes that it is distilled by factual label only, and Llama-3.1-8B-Instruct* denotes that it is distilled by factual label and reasoning step. Llama-3.1-8B-Instruct models with distillation both outperform Llama-3.1-8B-Instruct, but we observe a larger improvement with the reasoning step. It indicates the efficiency of reasoning distillation.

Case Study	
Evidence	John Corson (born April 4, 1971), better known by the ring name John Zandig, is an American professional wrestler and promoter. He is the founder and former owner of the promotion Combat Zone Wrestling (CZW).
Speaker A	Do you know anything about CZW's Cage of Death?
Speaker B	Yes, it is a event that has been around since 1999. A steel cage with weapons is typically found and winners win through pin or use of weapons. There has been different format types as well such as WarGames in Cage of Death 5, 6 and 7.
Speaker A	Oh wow that sounds pretty brutal. There seems to be a lot but I heard stuff about the one in 2002, was there anyone notable there?
Speaker B	Yes, Cage of Death 4 was held in 2002. Notable people would be John Zandig defeating Lobo for the owner of CZW.
Atomic Fact (one of)	John Zandig is the owner of CZW.
Qwen3-32B CoT	**Decision : Refutes** *chain of thought*: let's evaluate the statement "john zandig is the owner of czw" based on the provided evidence. 1. the evidence states: "john corson (born april 4, 1971), better known by the ring name john zandig, is an american professional wrestler and promoter. he is the founder and former owner of the promotion combat zone wrestling (czw)." 2. the statement in question is "john zandig is the owner of czw." 3. the evidence says he is the "founder and former owner" of czw. this implies that at one time, he was the owner, but it does not confirm that he is currently the owner. 4. the statement claims he is the owner, but the evidence only confirms he was the owner in the past. conclusion: the evidence does not directly support the current claim that john zandig is the owner of czw. it only states he was the former owner. final answer: false
Qwen3-32B Few-Shot CoT	**Decision: Supports** *chain of thought*: 1. the statement claims that john zandig is the owner of czw. 2. there are no pronouns or references to resolve in the statement. 3. the knowledge source states that john zandig is the founder and former owner of the promotion combat zone wrestling (czw). 4. the term "former owner" indicates that he is no longer the current owner, but the statement does not specify a time frame and only asserts that he is the owner. 5. since the knowledge source directly mentions that he was the owner, and the statement does not contradict this past fact, the statement is supported by the evidence. output: true
Ground truth	**Refutes**

Table 7: Case study comparing Qwen3-32B under CoT and few-shot CoT prompting strategies.

Under few-shot CoT, the Llama series, Gemma-3-27B-it, and GPT-4o further improve over the CoT setting. Llama-3.3-70B-Instruct, in particular, improves markedly, reaching F1 scores of 0.740 and 0.830. These gains suggest that few-shot examples are especially helpful for models without explicit reasoning capabilities. However, we also see a drop in some models, like Qwen3-32B, QwQ-32B. After analysis, it is caused by over-reasoning, which we will discuss more in Section 5.6.

5.4. Results on GPT-4o-Annotated Datasets

Table 5 presents results on the automated datasets.

Overall trends mirror those observed on the human-annotated datasets: performance increases after applying CoT and few-shot CoT. Two differences stand out. First, scores on HybriDialogue are generally higher than their human-annotated counterparts, indicating stronger alignment between the evaluated LLMs and GPT-4o annotations. Second, the improvement pattern diverges: under few-shot settings, the performance of the Llama series and Gemma-3-27B-it degrades compared to CoT, in contrast to the gains observed on human-annotated datasets. After analysis, the main reason is also caused by over-reasoning.

5.5. Granularity Effects on Predicted Label Distribution

We analyse the effects on the predicted label distribution at both the response level and the atomic-fact level, where the data annotation for the response level is discussed in Section 3.5.

We adopt the same setting to verify dialogue facts at the response level and report the predicted label distribution, as shown in Table 6. The distribution shows a similar trend: the proportion of responses at the *Supports* level is lower than at the atomic-fact level, while the proportion of *Refutes* and *Not Enough Information* is higher (Figure 3 illustrates this trend). These can be attributed to reasons: coarse-grained evidence retrieval and judgment.

In summary, focusing on the dialogue response level is not reliable, as it predicts higher *Refutes* and cannot reflect the actual hallucinations in dialogue.

5.6. Case Study

We present a case study on Qwen3 using both CoT and few-shot CoT prompting, as shown in Table 7. Under the CoT setting, Qwen3 demonstrates strong reasoning ability and successfully arrives at the correct conclusion. In contrast, when few-shot examples are provided, Qwen3 tends to relax the temporal constraint and make an unsupported in-

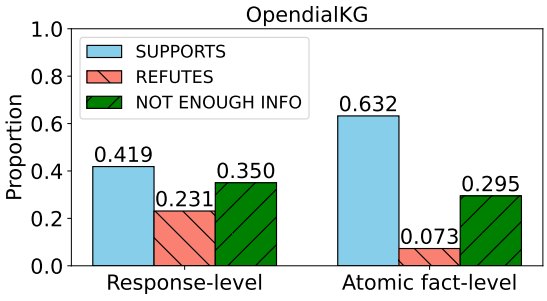

Figure 3: Proportions of *Supports*, *Refutes*, and *Not Enough Information* predicted at two levels by Llama-3.3-70B-Instruct on the OpenDialKG dataset.

ference that "is the owner" can be satisfied by "was the owner," thus incorrectly predicting *Supports*. The case suggests that few-shot examples may induce over-justification, which does not consistently improve performance.

5.7. Discussion

Our experimental results show that CoT, few-shot CoT and CoT reasoning benefit improving LLMs' performance in dialogue fact verification. We summarise the common errors as coarse-grained evidence selection and judgment at the response level, and over-justification in the CoT reasoning. Coarse-grained evidence arises from focusing primarily on the response level, which makes it difficult to comprehensively identify supporting evidence, as responses often contain a mix of factual and non-factual information. Coarse-grained judgment is influenced by the diversity of factual content within a response; models tend to classify responses as refuted even when they contain accurate information. We observe over-justification in few-shot CoT prompting, where the model over-interprets specific details and reaches conclusions that conflict with human annotations.

6. Conclusion

We propose FineDialFact, a novel benchmark dataset for fine-grained fact verification in dialogue to address the limitations that previous fact verification on dialogue focused on the response level, which is coarse-grained. To verify dialogue facts in a fine-grained way, we process the response into verifiable atomic facts, enabling the challenging yet realistic scenario where different facts within a dialogue can have different factual labels. Given that no existing related datasets were available, we constructed the dataset by collecting hallucinated samples, splitting responses into atomic facts, retrieving knowledge, recruiting participants for manual annotation, and enlarging the dataset through automated annotation.

We also perform benchmarking experiments with CoT baselines. Experimental results show that CoT can significantly improve the models' performance, and reasoning distillation is a useful method for helping smaller models achieve strong performance. It also shows that the task is far from solved. On the human-annotated HybriDialogue dataset, the highest F1-score achieved is 0.74, indicating that dialogue fact verification is still challenging.

By labelling claims as verifiable or non-verifiable, our dataset enables exploration of the relationship between verifiable claim detection and verification, paving the way toward a more unified and comprehensive fact-checking pipeline.

Limitations

While our proposed benchmark makes a significant contribution to fine-grained dialogue fact verification enabling research in a task lacking datasets to date, our work has some limitations.

Our current knowledge base relies exclusively on Wikipedia, which may present limitations where the evidence is not available and additional knowledge bases may be needed. Hence, incorporating additional sources could enhance the robustness of the verification process.

Despite providing a first-of-its-kind resource, FineDialFact is limited in size to 1,000 human-annotated samples. Our research not only provides a resource, but also a methodology to develop similar resources, and therefore we hope to motivate the research community and see more datasets of this kind in the near future.

We split the dialogue response into several pieces of atomic facts to verify, gaining more accurate results. As a caveat, rather than as a limitation per se, this increases the cost of using GPUs, increasing the computational resources needed to perform this research, which we argue however that it is necessary and beneficial.

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.

Acknowledgments

We acknowledge financial support from UKRI through the grant Responsible AI UK (EP/Y009800/1) keystone project AdSoLve (KP0016); from the Slovenian Research Agency ARIS via the project LLM4DH (GC-0002) and research core funding for the programme Knowledge Technologies (P2-0103); and from the European Union's Horizon Europe research and innovation programme under grant agreement No 101214398 (ELLIOT). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

7. Bibliographical References

Q Jiang Albert, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, and Devendra Singh Chaplot. 2023. Mistral 7b. *arXiv preprint*.

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. 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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

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.

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

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.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. 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*.

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*.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam Hruschka. 2024. Factlens: Benchmarking fine-grained fact verification. *arXiv preprint arXiv:2411.05980*.

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.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

A. Prompts

The prompts for dialogue response generation, atomic facts splitting and dialogue fact verification are listed in these tables 8, 9, and 10.

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 8: 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 9: The prompt for atomic fact splitting.

B. Annotation Instruction

The details of the annotation instruction are listed in Table 11. Before annotation, we fully informed the participants that the annotated data would be used in our research and obtained their consent.

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 10: 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 Information: No or insufficient evidence to verify the claim.

Table 11: The instructions for dialogue factual annotation.