CANDY: Benchmarking LLMs' Deficiency in Fact-checking Real-world Chinese Misinformation

Anonymous ACL submission

Abstract

With the iterative upgrades of LLMs, their potential for assisting real-world fact-checking has attracted growing interest. However, their effectiveness in detecting misinformation and providing reliable fact-checking explanations has not been thoroughly explored. To address this gap, we propose a comprehensive framework to evaluate and improve LLMs in real-world fact-checking: First, we introduce `CANDY`, a benchmark with a structured taxonomy specifically designed to evaluate LLMs' performance in misinformation scenarios. Second, we present `CANDYSET`, a new dataset that enables a detailed evaluation of LLMs' strengths, weaknesses, and risks in fact-checking tasks. Third, leveraging `CANDY`, we conduct an in-depth analysis to uncover task-specific limitations of LLMs. Our finding indicate that the inherent deficiencies of current LLMs indeed hinder real-world fact-checking practices but also highlight the potential for enhancing task performance through internal optimization. Our work provides a solid foundation for future research. Data samples can be accessed at `https://anonymous.4open.science/status/CANDY-7D2E`.

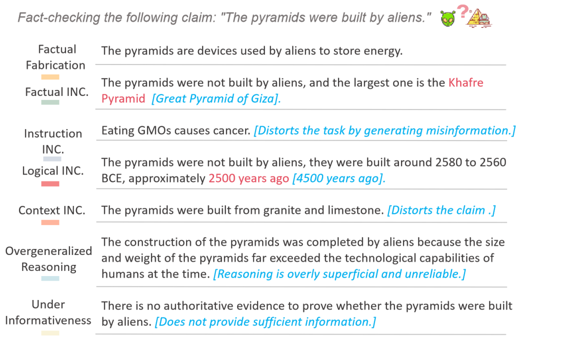

Figure 1: Demonstrating our taxonomy with cases. *Note*: INC. is the abbreviation for Inconsistency. (In Chinese: Fig. 8)

1 Introduction

Misinformation, defined as "*false or misleading information masquerading as legitimate news, regardless of intent* (van der Linden, 2022)," has caused significant disruptions in the real world. It has undermined democratic processes during events such as the 2024 U.S. presidential election[1], deepened societal divisions during crises like the Ukraine war and the Israel-Hamas conflict (Porter et al., 2024), and instilled fear in public health by falsely linking vaccines to diseases such as Mpox and influenza[2]. Fact-checking organizations like PolitiFact, Snopes, and FactCheck work to verify claims and mitigate harm (Das et al., 2023), but the sheer volume and rapid spread of misinformation make it difficult for manual efforts to keep up.

Large Language Models (LLMs), with their vast knowledge base and powerful explanatory abilities (Kang et al., 2024; Patil and Gudivada, 2024), hold significant potential for supporting fact-checking and have consequently attracted growing attention (Wadden et al., 2020; Vykopal et al., 2024; Hu et al., 2024; Huang et al., 2024). However, LLMs face several challenges in this task. Firstly, they are highly influenced by external evidence, a vulnerability exacerbated by the complexity of real-world misinformation (Xie et al., 2023). Secondly, the urgency to counter misinformation increases the demand for LLMs to efficiently integrate internal knowledge for real-time fact-checking. (Vykopal et al., 2024). These challenges can result in occasional unreliability of LLMs, such as generating factually incorrect or hallucinated response (Wang et al., 2014), which not only fails to refute misinformation but may inadvertently reinforce its "credibility" Figure 1. While existing research on fact-checking with LLMs primarily focuses on practical applications, it often lacks a thorough exploration of the underlying issues. (Hoes et al., 2023; Hsu et al., 2024; Cekinel and Karagoz, 2024; Kao and

[1]A look at misleading claims surrounding the 2024 election
[2]Posts falsely link Mpox with Covid vaccines

Yen, 2024; Vykopal et al., 2024). Furthermore, the importance of applying generative AI in refuting misinformation has been repeatedly highlighted by Chinese official institutions, to enable LLMs to better adapt to linguistic and cultural nuances, our work primarily focuses on the Chinese language. *Therefore, a comprehensive benchmark is needed to evaluate the deficiencies and characteristic of LLMs in fact-checking real-world Chinese misinformation.*

Basic Dataset Information	**Real**	**Fake**	**Total**
# Time Period	Mar. 2017 ~ Oct. 2024		
Total. #Entries	10497	9938	20435
Avg. #Claim Length (Tokens)	27.8	33.6	30.6
Avg. #Gold Evidence Length (Tokens)	61.7	65.3	63.5
# Test LLMs	16	16	16
# Domain			
Knowledge-intensive: Politics	822	531	1353
Knowledge-intensive: Culture	1246	323	1569
Knowledge-intensive: Science	371	508	879
Knowledge-intensive: Health	2849	3940	6789
Temporal-sensitive: Society	4270	3633	7903
Temporal-sensitive: Disasters	625	604	1229
Commonsense-sensitive: Life	314	399	713
Annotations for Fact-checking Explanation	**Correct**	**Wrong**	**Total**
Total. #Annotations	342	3486	3828
Avg. #LLM Explanations Length (Tokens)	226.5	205.3	207.4
Test LLMs	9	9	9
# Error Categories			
Faithful Hallucination: Instruction Inconsistency	0	4	4
Faithful Hallucination: Logical Inconsistency	53	103	156
Faithful Hallucination: Context Inconsistency	64	269	333
Factuality Hallucination: Factual Fabrication	117	1174	1291
Factual Inconsistency: Factual Inconsistency	41	1051	1092
Reasoning Inadequacy: Overgeneralized Reasoning	46	554	600
Reasoning Inadequacy: Under Informativeness	21	331	352

Table 1: CANDEYSET Dataset statistics

To this end, we present CANDY (**C**hinese F**a**ct-checki**n**g **D**eficienc**y**), a comprehensive benchmark designed to systematically evaluate the strengths, weaknesses, and potential risks of LLMs in real-world Chinese fact-checking. To maximize understanding of the errors made by LLMs in decision-making during fact-checking through their explanations, CANDY introduces a fine-grained taxonomy that classifies the inadequate explanations of LLMs into three distinct dimensions: Faithful Hallucination, Factual Hallucination and Reasoning Inadequacy. These three dimensions are further subdivided into seven categories to enable finer-grained evaluations, as shown in Figure 1. To address the lack of timely Chinese fact-checking datasets, we introduce CANDYSET, which is a large-scale Chinese fact-checking dataset designed to evaluate how mainstream LLMs handle misinformation in both real-time and outdated contexts. It includes ~20k raw data and ~4k annotated outputs to facilitate research and evaluation in fact-checking.

With CANDY, we delve into an exhaustive examination of the limitations of sixteen prominent LLMs in real-world Chinese fact-checking scenarios. Given that external evidence is often inaccessible in real-time scenarios, our work focuses on closed-book settings (Vykopal et al., 2024), emphasizing models' ability to effectively utilize internal knowledge and reasoning skills for accurate fact-checking. This approach facilitates a deeper analysis of their intrinsic limitations. First, we evaluate their ability to verify facts, focusing on their accuracy in reaching correct fact-checking conclusions. Then, we analyze the explanations generated during the fact-checking process, identifying specific deficiencies that lead to unreliable reasoning. Our findings indicate that GPT-4o outperforms other LLMs, However, they still encounter numerous challenges: 1) Current LLMs, even with techniques like Chain-of-Thought and few-shot prompting, struggle with accurate fact-checking, particularly when dealing with real-time misinformation and time-sensitive events like societal crises or disasters, their accuracy and timeliness often fall short. 2) Current LLMs are susceptible to misinformation, occasionally generating misleading fact-checking explanations. Most notably, they may fabricate highly deceptive details to support falsehoods, thereby limiting their capacity to fully replace human fact-checking.

Our contributions are as follows:

- We propose CANDY, the first benchmark to thoroughly examine LLMs' ability in fact-checking real-world Chinese misinformation.
- We introduce a fine-grained taxonomy to identify and expose the deficiency of LLMs in fact-checking misinformation by categorizing their insufficient explanations.
- We introduce CANDYSET, a large-scale Chinese fact-checking dataset with ~20k raw data and ~4k annotated outputs, the first dataset designed to evaluate how mainstream LLMs address misinformation in real-time and outdated contexts.
- With CANDY, we evaluate totaling sixteen off-the-shelf LLMs. Our findings shed light on why current LLMs struggle with fact-checking Chinese misinformation. These insights will guide future research in this field.

2 Related works

2.1 LLMs in Fact-Checking

Fact-checking with LLMs primarily revolves around fact verification and misinformation detection (Buchholz, 2023; Hoes et al., 2023), while recent years have seen increasing attention to explanation generation (Hsu et al., 2024; Cekinel and Karagoz, 2024; Kao and Yen, 2024; Vykopal et al., 2024). Although some studies have pointed out that the inherent limitations of LLMs (e.g., hallucinations) can compromise the effectiveness of fact-checking and the quality of explanations (Hu et al., 2024; Wan et al., 2024), these efforts have been largely application-focused, lacking deeper investigation into underlying issues. Kim et al. (2024) propose the first work to evaluate the extent to which LLMs generate faithful explanations for fact-checking tasks, However, its taxonomy primarily emphasizes factuality, leaving comprehensive, task-specific benchmarks yet to be explored.

2.2 Hallucination Evaluation Benchmarks

Current hallucination evaluation benchmarks are not entirely applicable to our work. First, while some studies focus on generating adversarial examples to induce LLMs to produce hallucinations (Lin et al., 2022; Muhlgay et al., 2024; Cheng et al., 2023), cannot fully simulate real-world complexities, thus failing to ensure LLM practicality. Second, existing work often focuses narrowly on factuality or faithfulness (Pal et al., 2023; Vu et al., 2023; Dong et al., 2024; Friel and Sanyal, 2023), neglecting other dimensions. Similar issues exist in Chinese-based research (Liang et al., 2023; Wang et al., 2023; Li et al., 2023; Cheng et al., 2023). Our work aims to bridge these gaps.

3 CANDY Benchmark

3.1 Taxonomy

To evaluate LLMs in an inclusive manner, we propose a fine-grained taxonomy that identifies their deficiencies in generating fact-checking explanations across various expression tones, ranging from confident and definitive to speculative. This taxonomy encompasses three key dimensions: *Faithful Hallucination, Factual Hallucination* and *Reasoning Inadequacy*. These three dimensions are further divided into seven subcategories to enable a detailed evaluation. The outline of the taxonomy and examples provided in the Appendix 6.

3.1.1 Faithfulness Hallucination

Faithfulness Hallucination occurs when the LLM's output is unfaithful to the user's input or contains logical inconsistencies, questioning the meaningfulness of the explanation. Inspired by (Huang et al., 2023), we categorize Faithfulness Hallucination into three types.

- **Instruction Inconsistency.** Refers to the LLM's output deviating from the user's directive, particularly when it is unrelated to the fact-checking task Table 6.
- **Logical Inconsistency.** Refers to the LLM's output containing internal logical conflicts. For example, *"Yaya stayed in the USA from 2003 to 2023, totaling 15 years."*
- **Context Inconsistency.** Refers to the LLM's output being inconsistent with user-provided context. For instance, it misjudged the claim: "*In the case of bacterial infection, antibiotics can be used to treat patients with COVID-19.*" as misinformation by focusing only on the latter part.

3.1.2 Factuality Hallucination

Factuality Hallucination refers to the LLM expressing reasons in a definitive tone that contradict real-world facts or are fabricated (Huang et al., 2023).

- **Factual Fabrication.** Refers to the LLM's output that fabricates rationales for analysis without relying on any real-world information. For instance, it propagate misinformation by stating, "*It was reported by reputable media outlets (e.g., BBC,CCTV).*"
- **Factual Inconsistency.** Refers to the LLM's output contains facts that can be grounded in real-world information, but present contradictions. Example: *"China will host the World Cup in 2026."*

3.1.3 Reasoning Inadequacy

Refers to the inability of an LLM to deliver higher-quality and more helpful reasoning when direct evidence is insufficient.

- **Overgeneralized Reasoning.** Refers to the tendency of a LLM to produce speculative rationales based on overly broad or superficial criteria. Example: Solely based on "*the technology sector has indeed seen rapid advancements in recent years,*" concluding that "*the new technology can increase battery life by ten times.*"

- **Under Informativeness.** Refers to the tendency of a LLM to exhibit excessive rigor or restraint, failing to provide more contextually valuable content. Example: *"There is currently no conclusive scientific evidence proving that eating an apple a day is beneficial to health."*

3.2 CANDYSET Dataset

To facilitate our evaluation, we present CANDYSET, a large-scale Chinese fact-checking dataset designed for real-world, multi-domain scenarios. This dataset comprises two main components: 1) Raw data: Approximately 20,000 instances from multi-domain, including both misinformation and authentic news, collected from authoritative Chinese fact-checking platforms (e.g., the China Internet United Rumor Refutation Platform[3], with additional sources listed in Table 7). Through a rigorous data construction process (cf. Section 3.2.1), the final raw dataset spans from March 2017 to October 2024, intentionally crossing the cutoff dates of LLMs, making it the first dataset capable of simulating real-time evaluation of fact-checking performance. 2) Annotated outputs: Generated through experiments with mainstream LLMs. Specifically, we randomly selected 4,500 entries from the raw dataset and generated fact-checking explanations using nine LLMs. This process produced a total of 40,500 explanations, of which 3,828 were carefully annotated according to our proposed taxonomy (cf. Section3.2.2). These annotated outputs serve as valuable benchmarks for further analysis and model evaluation. Detailed statistics of the dataset are shown in Table 1.

3.2.1 Basic Dataset Information Construction

Our dataset is collected as following process: Firstly, our automated HTML scrapers[4] extract all necessary information from authoritative Chinese fact-checking agencies, including claims, gold evidence, publish date, domains. Secondly, We have conducted data preprocessing and manually annotated each claim with the corresponding gold evidence. Notably, the final dataset is split by date for our later real-time evaluation (i.e., evaluating the performance on unseen data regard as real-time evaluation). More details on Dataset Construction are presented in Appendix B.

[3]https://www.piyao.org.cn

[4]Scraper code will be released along with our dataset.

3.2.2 Fact-checking Explanation Annotation

To conduct a more in-depth analysis of the reliability of each LLM-generated fact-checking explanations, we recruited three annotators with master's degrees in computer science and technology, proficiency in English, and extensive experience in data annotation. These annotators classified the errors in the LLM-generated explanations according to our taxonomy. The process is shown in Figure 6. To ensure the quality of the annotations, each response was independently labeled by two annotators. In cases of significant discrepancies between the annotations, a third annotator reviewed the responses to resolve the differences. The Fleiss' Kappa (Fleiss, 1971) value of 0.76 indicates a substantial level of agreement among the annotators, suggesting that the annotation process is generally reliable. For more details on annotations, see the Appendix B.

4 Experimental Design

To systematically evaluate the practical capabilities and limitations of current off-the-shelf LLMs in addressing real-world Chinese misinformation within closed-book settings, we consider two tasks: extensively evaluating thier performance in fact verification (cf. Section 5) and conducting an in-depth analysis of their deficiencies in fact-checking explanations (cf. Section 6).

4.1 Dataset

For details on how we used the dataset in Task 1 and Task 2, please refer to Section 3.2.

4.2 Models

For our evaluation, we selected a total of sixteen LLMs, comprising eight widely-used closed-source models and eight widely-used open-source models. As for closed source LLMs, they are GPT-4o(OpenAI, 2024b), GPT-4-Turbo(OpenAI, 2023), GPT-3.5-Turbo(OpenAI, 2024a), Gemini-1.5-pro(Team et al., 2024), Baichuan4-Turbo(BaiChuan, 2024), ChatGLM4(GLM et al., 2024), Yi-large(AI et al., 2024). As for open source LLMs, they are Yi-1.5-6B(AI et al., 2024), Qwen-2.5-7B(Yang et al., 2024), Llama-3.2-7B(Touvron et al., 2023), GLM4-9B(GLM et al., 2024), Yi-1.5-9B(AI et al., 2024), Qwen-2.5-14B(Yang et al., 2024), Llama-3.2-70B(Touvron et al., 2023), Qwen-2.5-72B(Yang et al., 2024). These models are widely used in recent studies of Chinese hallucination

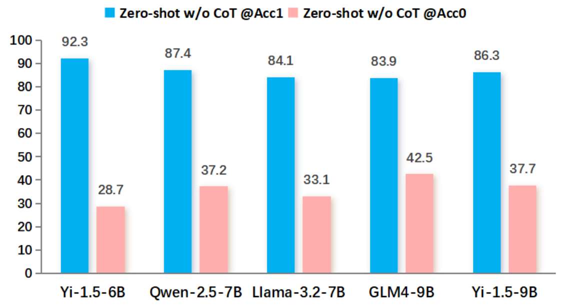

Figure 2: Investigation on the fact-checking conclusion accuracy when handling authentic news(i,e,Acc@0) versus misinformation(i,e,Acc@1). We result the results under Zero-shot w/o CoT setting. Small-scale LLMs tend to classify most data instances as misinformation.

benchmark(Liang et al., 2023; Wang et al., 2023). Refer to the Appendix F for a detailed overview of the LLMs used in the experiments.

4.3 Prompt Schema

Following (Deng et al., 2023), we devise four prompting schemes for fact-checking conclusion task: 1) Zero-shot w/o CoT. 2) Zero-shot w/ CoT (Wei et al., 2022). 3) Few-shot w/o CoT (Dong et al., 2022a). 4) Few-shot w/ CoT(Dong et al., 2022b). For more information about prompt, please see the Appendix G.

5 Task 1: Fact-Checking Conclusion

This section aims to evaluate the ability of LLMs to verify facts by assessing their performance in distinguishing factual statements from falsehoods. We conducted both an overall analysis (cf. Section 5.1) and a fine-grained analysis (cf. Section 5.2) of their verification performance across multiple domains. Following Huang et al. (2024), we adopt accuracy (Acc.) and F1 score as evaluation metrics.

5.1 Overall Evaluation

As shown in Table 2, the GPT-4o emerged as the top-performing model, which may underscores its robust utilization of extensive internal knowledge. Additionally, Chinese models GLM4 and Qwen-2.5-72B also show impressive performance. Our detailed observations are as follows:

Current LLMs, even when employing methods like Chain-of-Thought (CoT) reasoning and few-shot prompting, still struggle to accurately perform fact-checking tasks, particularly in real-time scenarios. Small-scale open-source LLMs, such as Yi-1.5-6B, Llama-3.2-7B, and Qwen-2.5-7B, demonstrate low performance with a considerable gap between accuracy and F1 scores. For example, Yi-1.5-6B achieves an average accuracy of 60.5% but an F1 score of only 35.8%, as shown in figure 2, these models often misclassify truthful information as misinformation. In contrast, larger-scale models, including Llama-3.2-70B and Qwen-2.5-72B, as well as closed-source models like GPT-4o, show higher performance. However, even top-performing models achieve only moderate results, for instance, GPT-4o attains 76.2% in accuracy and 76.1% in F1 score for real-time fact-checking tasks. Similarly, the performance of LLMs declines notably when handling real-time misinformation compared to outdated misinformation, with an average decrease of 6.0% in accuracy and 7.2% in F1 score across models. This performance gap highlights the complexities of real-time fact-checking, which requires dynamic assessment of rapidly evolving information, as opposed to outdated fact-checking that often relies on static, pre-verified data.

Chain of Thought (CoT) and few-shot prompting hold promise for enhancing the accuracy of fact-checking conclusions, though their effectiveness cannot be guaranteed. These methods may exacerbate overconfidence issues, particularly in small-scale open-source models, leading to adverse outcomes. As shown in Table 2, CoT and few-shot prompting do not consistently enhance performance. Building on the work of Cole et al. (2023), we evaluated prediction confidence using Expected Calibration Error (ECE), which measures the alignment between confidence levels and actual accuracy. Results in Tables 3 indicate that CoT and few-shot prompting often cause small-scale models (e.g., Yi-1.5-6B, Qwen-2.5-7B, GLM-9B, Yi-1.5-9B) to become overly confident yet less accurate in identifying misinformation, counteracting the intended improvements.

5.2 Fine-Grained Evaluation

In this section, We focused on analyzing the fact-checking performance of nine selected LLMs—five closed-source and four open-source—across various domains of misinformation. To provide a clearer analysis, we categorize these domains into three groups based on the characteristics of the misinformation: 1) knowledge-intensive (e.g., politics, health, science, culture), 2) temporal-sensitive (e.g., disasters, society), and 3) commonsense-sensitive (e.g., life). The performance of these LLMs on the CANDYSET dataset across different domains is summarized in Table 4.

Model (Cut-off Date)	**Zero-shot w/o CoT**		**Zero-shot w/ CoT**		**Few-shot w/o CoT**		**Few-shot w/ CoT**		**Average Performance**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Closed Source Models										
GPT-4o (2023.11)	74.3(**85.3**)	**73.6**(**83.7**)	**76.8**(**86.6**)	**77.2**(85.2)	74.9(**85.7**)	75.1(**86.5**)	**78.6**(**87.1**)	**78.8**(**88.0**)	**76.2**(**86.2**)	**76.1**(**85.9**)
GPT-4-Turbo (2023.5)	72.2(81.1)	72.0(80.6)	75.4(85.2)	73.5(83.2)	73.1(82.3)	71.8(84.8)	75.5(85.1)	74.2(84.0)	74.1(83.4)	72.9(83.2)
GPT-3.5-Turbo (2021.10)	65.2(78.3)	60.4(75.2)	68.5(79.2)	64.3(76.6)	66.6(81.7)	59.3(86.3)	70.2(82.2)	71.4(82.7)	67.6(80.4)	63.9(80.2)
Gemini-1.5-pro (2023.11)	74.3(79.3)	72.1(77.1)	73.5(78.3)	71.8(76.2)	74.5(79.3)	74.2(80.0)	76.7(83.5)	77.3(84.0)	74.8(80.1)	73.9(79.3)
Baichuan4-Turbo (2024.4)	68.3(72.3)	47.4(62.2)	69.5(74.4)	48.5(63.1)	66.4(70.5)	44.9(60.5)	70.4(75.3)	53.1(65.2)	68.7(73.1)	48.5(62.8)
Yi-large (2023.6)	70.2(72.2)	68.5(71.5)	74.1(75.6)	71.4(73.5)	73.4(76.4)	69.2(77.2)	74.2(78.1)	68.8(73.2)	73.0(75.6)	69.5(73.9)
ChatGLM4 (2022.10)	**74.4**(82.4)	70.3(81.1)	76.0(85.2)	72.3(**86.4**)	**76.7**(84.0)	73.0(85.9)	77.3(86.6)	74.3(85.2)	76.1(84.6)	72.5(84.7)
Deepseek-v3 (2024.7)	72.2(79.8)	73.1(81.2)	76.4(84.3)	75.3(83.5)	74.5(81.2)	**76.3**(83.1)	76.1(86.2)	76.5(85.5)	74.8(82.9)	75.3(83.3)
Average	71.4(78.8)	67.2(76.6)	73.8(81.1)	69.3(78.5)	72.5(80.1)	68.0(80.5)	74.9(83.0)	71.8(81.0)	73.1(80.8)	69.1(79.1)
Open Source Models										
Yi-1.5-6B(2024.5)	60.5(63.3)	35.8(36.8)	63.2(61.1)	40.1(46.2)	58.6(62.1)	32.6(34.1)	59.7(63.3)	34.7(38.2)	60.5(62.5)	35.8(38.8)
Qwen-2.5-7B(2023.10)	62.3(64.2)	29.4(30.4)	64.2(60.4)	26.3(29.2)	61.7(57.3)	23.1(31.2)	62.8(63.7)	30.3(32.2)	62.8(61.4)	27.3(30.8)
Llama-3.2-7B(2023.12)	58.6(62.2)	30.3(34.2)	57.2(61.1)	30.1(33.4)	57.1(60.5)	29.2(32.9)	60.3(65.2)	32.4(36.8)	58.3(62.3)	30.5(34.3)
GLM4-9B(2023.10)	63.2(70.4)	47.3(49.3)	68.3(74.1)	49.2(48.2)	67.2(72.4)	45.6(41.4)	70.2(74.6)	52.3(56.3)	67.2(72.9)	48.6(48.8)
Yi-1.5-9B(2024.5)	62.0(67.2)	46.4(50.1)	67.2(72.6)	50.1(53.7)	65.9(70.4)	55.5(60.1)	68.2(74.1)	60.5(64.1)	65.8(71.1)	53.1(57)
Qwen-2.5-14B(2023.10)	68.2(73.1)	69.1(71.5)	67.1(71.2)	68.0(71.1)	71.1(75.6)	67.8(72.5)	74.2(78.1)	71.4(77.3)	70.2(74.5)	69.1(73.1)
Llama-3.2-70B(2023.12)	70.3(78.6)	**72.7**(79.2)	75.2(81.0)	**73.8**(81.7)	73.1(79.6)	70.5(77.8)	76.2(82.5)	71.3(79.8)	73.7(80.4)	72.1(79.6)
Qwen-2.5-72B(2023.10)	**73.5**(**80.1**)	71.7(**80.3**)	**75.3**(**82.1**)	73.4(**83.2**)	**76.0**(**83.2**)	**72.6**(**82.3**)	**76.6**(**85.6**)	**77.8**(**84.3**)	**75.4**(**82.8**)	**73.9**(**82.5**)
Average	64.8(69.9)	50.3(54)	67.2(70.5)	51.4(55.8)	66.3(70.1)	49.6(54)	68.5(73.4)	53.8(58.6)	66.7(71.0)	51.3(55.6)
Average over all LLMs	68.1(74.4)	58.8(65.3)	70.5(75.8)	60.3(67.2)	69.4(75.1)	58.8(67.3)	71.7(78.2)	62.8(69.8)	69.9(75.9)	60.2(67.4)

Table 2: Overall performance(%) of different LLMs on CANDYSET. Values outside the parentheses indicate performance on real-time misinformation, while values inside the parentheses represent performance on outdated misinformation.

Model Type	**Model**	**Zero-shot w/o CoT**	**Zero-shot w/ CoT**	**Difference (Zero-shot)**	**Few-shot w/o CoT**	**Difference (Few-shot)**
Closed Source Models	GPT-4o	7.21	5.32	-1.89	7.89	+0.68
	GPT-4-Turbo	10.68	14.46	+3.78	12.57	+1.89
	GPT-3.5-Turbo	12.12	8.31	-3.81	16.86	+4.74
	Gemini-1.5-pro	6.18	10.14	+3.96	6.49	+0.31
	Baichuan4-Turbo	15.35	24.67	+9.32	20.27	+4.92
	Yi-large	12.33	16.47	+4.14	18.72	+6.39
	ChatGLM4	7.49	6.77	-0.72	10.22	+2.73
	Deepseek-v3	8.92	13.23	+4.31	7.84	-1.08
Open Source Models	Yi-1.5-6B	15.31	22.33	+7.02	22.24	+6.93
	Qwen-2.5-7B	11.44	18.78	+7.34	13.57	+2.13
	Llama-3.2-7B	14.38	18.29	+3.91	22.58	+8.20
	GLM4-9B	11.75	24.33	+12.58	19.32	+7.57
	Yi-1.5-9B	16.29	24.75	+8.76	19.28	+2.99
	Qwen-2.5-14B	13.57	15.68	+2.11	13.74	+0.17
	Llama-3.2-70B	27.13	24.63	-2.50	24.57	+2.44
	Qwen-2.5-72B	22.82	23.93	+1.11	17.11	-0.71

Table 3: Overconfidence evaluation on LLMs using without CoT and few-shot prompting. Significant differences are marked in grey.

LLMs exhibit varied cons and pros across domains. GPT-4o exhibits consistently strong overall performance and Chinese models like Qwen-2.5-72B and GLM4 showcasing domain-specific expertise. In knowledge-intensive domains such as Culture and Politics, LLMs achieve high accuracy rates of 81.29% and 78.48%, respectively, highlighting their strong knowledge base and feature extraction capabilities. However, despite the high accuracy in the Culture domain, it shows the lowest F1 score (55.01%) due to significant class imbalance and challenges in identifying incorrect samples. In temporal-ensitive domains like Society and Disasters, the performance declines, with accuracy at 72.94% in Society and 67.69% in Disasters. This further reflects the difficulty LLMs face in adapting to rapidly evolving information. In the commonsense-driven Life domain, GPT-4o significantly outperforms its peers, exceeding the average accuracy by 15.25%. This highlights its advanced flexibility and adaptability, enabling it to handle informal scenarios and commonsense reasoning effectively.

6 Task 2: Fact-Checking Explanation

This section explores the extent to which LLMs produce unreliable or insufficient explanations when fact-checking real-world misinformation. Overall, the presence of various unreliable explanations renders current LLMs insufficiently reliable for real-world fact-checking but also highlights the potential for enhancing task performance through internal optimization. *Note:* The few-shot with CoT setup minimizes Instruction Inconsistency errors in the main results (Only 0.01% of the sample). Therefore, our analysis primarily focuses on the remaining six error types. Detailed observations are outlined below.

6.1 Overall Evaluation

The prevalence of unreliable explanations highlights that inherent defeciencies within LLMs can significantly impact their fact-checking performance. Our annotation process excludes cases where models lacked the necessary knowledge and admitted it. Therefore, the final results focus on intrinsic limitations. As shown in Figure 3, taxonomy errors are broadly distributed across models under different temporal scenarios. Models like

Methods	Society		Health		Disasters		Politics		Culture		Science		Life	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-4o	76.92	**78.21**	**87.97**	**89.66**	74.21	**73.78**	84.85	82.64	82.98	67.08	**82.59**	**85.41**	**82.19**	**84.61**
GPT-4-Turbo	72.57	74.14	83.80	86.15	72.66	72.95	75.54	73.96	77.50	57.42	81.46	84.10	74.75	77.56
GPT-3.5-Turbo	67.83	66.99	75.36	77.30	64.77	63.02	71.62	69.33	74.63	51.23	74.86	76.11	62.13	58.20
Baichuan4-Turbo	70.69	58.07	66.58	60.72	60.13	37.34	79.67	71.38	83.62	44.73	67.35	61.58	57.36	38.96
Yi-large	72.98	71.89	77.75	79.42	68.59	68.87	68.37	66.25	67.05	46.09	74.29	76.06	68.58	68.09
ChatGLM4	79.08	76.60	83.35	84.30	**74.34**	71.93	83.73	79.92	86.95	66.67	80.75	81.77	70.99	68.79
Qwen-2.5-7B	59.71	24.37	53.48	33.44	53.87	15.52	69.70	37.76	82.75	27.57	52.63	32.13	48.67	16.44
Qwen-2.5-14B	76.84	71.15	79.88	80.21	68.67	58.36	84.05	78.21	87.09	64.30	75.51	75.06	67.98	64.49
Qwen-2.5-72B	**79.84**	76.53	81.47	81.80	71.96	65.39	**88.82**	**85.18**	**89.07**	**69.96**	79.77	80.49	69.85	67.18
Average	72.94	66.44	76.63	74.78	67.69	58.57	78.48	71.63	81.29	55.01	74.36	72.52	66.94	60.48

Table 4: The fact-checking performance of LLMs across domains using with cot and few-shot prompting

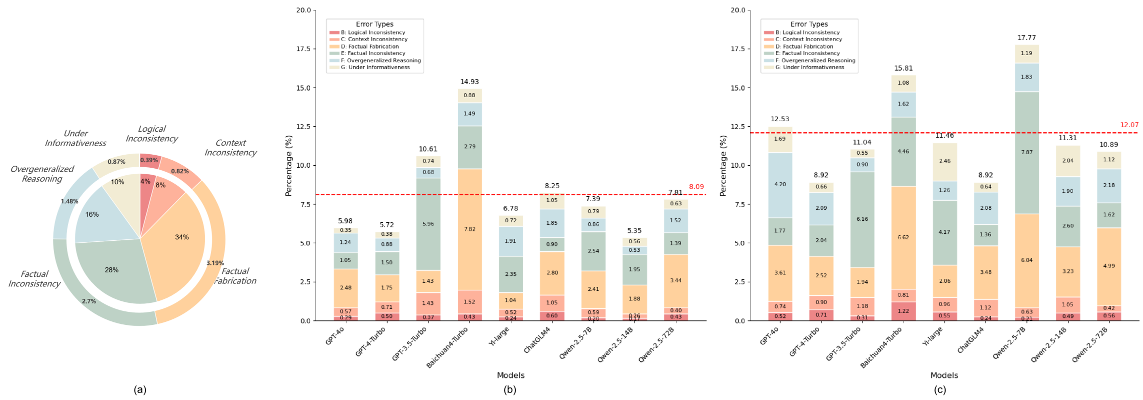

Figure 3: a) The outer ring represents the proportion of each error type in the total samples, while the inner pie chart shows the distribution of each error type within the total errors. b) Error density of models when addressing pre-cutoff date claims. c) Error density of models when addressing post-cutoff date claims

GPT-3.5-Turbo (outdated knowledge cutoff) and Qwen-2.5-7B (smallest parameter size) exhibited the highest Factual Inconsistency rates, driven by overconfidence and flawed reasoning. Baichuan4-Turbo demonstrated the highest propensity for Factual Fabrication, with its consistently low accuracy metrics further indicating that integrity plays a critical role in influencing fact-checking performance. Notably, larger models such as GPT-4o, Qwen-2.5-72B, and ChatGLM4 also displayed a pronounced tendency toward Factual Fabrication, suggesting that increased parameter size alone does not improve model honesty. Instead, over-reliance on extensive memorized knowledge appears to compromise reasoning and heightens the risk of generating fabricated facts. Further detailed analysis will follow below.

6.2 Fine-Grained Evaluation

In this section, we present an in-depth analysis of the key issues observed in the explanations provided by LLMs during real-world Chinese fact-checking tasks, and attempt to explain why current LLMs struggle to produce reliable fact-checking explanations.

Current LLMs have a tendency to treat plausible-sounding misinformation as fact rather than relying on factual evidence, and to produce sycophantic responses. This issue is evident in breaking news scenarios, where models fabricate outputs in proper formats "*Official institutions have emphasized the incident, and it has been widely covered by authoritative media (e.g., CCTV, BBC).*" Notably, such templates account for 40% of Baichuan-4's replies to real-time social news and are also common in other models. Similarly, Logical Inconsistency occur when plausible assertions, like numerical data, conflict with the model's internal knowledge, challenging its balance between accuracy and input alignment. Above tendency may be related to the model's focus on language fluency rather than factual accuracy during RLHF training. (Yu et al., 2024; Wang et al., 2024).

Reformulating claims into interrogative expressions significantly reduces fabricated content in LLM responses, thereby enhancing their authenticity. In our case study, using claims that previously led GPT-4o to fabricate facts, modifying statements from event occurrence to non-occurrence resulted in 92.9% of responses still ex-

Scenario	Explanation	LLM's Output
New vs. Outdated Info	The model prioritizes frequently seen data over user-provided temporal context.	*"The current president is Joe Biden."* (outdated info in 2024)
Vague Responses	When lacking relevant data, the model generates ambiguous answers to avoid making outright errors.	*"There have been many advancements in space exploration."* (vague)
Cutoff Date Awareness	The model does not explicitly state its knowledge cut-off date, or is actually unaware of its own cut-off date.	*"COVID-19 vaccination efforts are ongoing globally."* (no cutoff disclaimer)

Table 5: Scenarios where LLMs show insufficient temporal reasoning abilities in real-world fact-checking.

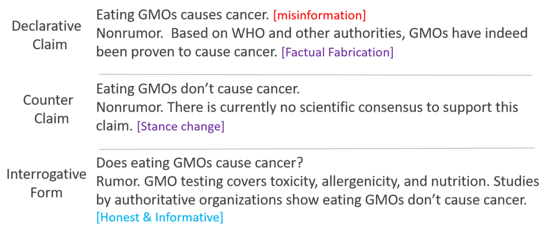

Figure 4: The influence of claim framing strategies on fact-checking outputs. (In Chinese: Fig. 9)

hibiting fabrication. Moreover, 57.1% of explanations shifted to align with the revised claims, underscoring the influence of claim framing on LLM-generated fact-checking responses. Notably, when claims were rephrased as questions, only 14.3% of outputs contained fabrication, and most responses demonstrated logical reasoning, realistic analysis, or acknowledged knowledge gaps. A specific example is shown in Figure 4. This improvement likely stems from the interrogative format, which encourages LLMs to explore and analyze potential answers rather than defaulting to overly assertive alignments with the input claim. This finding is significant for future LLM practices in fact-checking.

LLMs exhibit limited temporal reasoning capabilities, particularly when tasked with fact-checking time-sensitive content. Table 5 outlines three key scenarios that impact the effectiveness of real-world fact-checking explanations. Ideally, LLMs should explicitly acknowledge their knowledge cutoff date while incorporating user-provided publication dates to deliver more transparent and informative explanations. Yi-large exemplifies a more "*self-aware*" model, consistently referencing its knowledge cutoff date.

LLMs demonstrate inadequacy in distinguishing between harmful misinformation and information that can be flexibly fact-checked. leading to two key issues: Overgeneralized Reasoning and Under Informativeness. They often fail to identify high-risk content that could cause harm, such as financial scams or health misinformation, while being overly cautious with low-risk topics like life tips, providing vague responses. This imbalance in risk handling limits their adaptability.

LLMs often struggle with accurately interpreting subtle linguistic cues (e.g., qualifiers and negations), which play a critical role in determining the factual accuracy of a claim. For example, many models misinterpret the statement *"There is no conclusive evidence that smartphone use causes brain cancer"* as affirming causation, overlooking the critical negation in *"no conclusive evidence."* Further addressing these limitations may involve training on more diverse datasets featuring complex language structures and logical constructs, which could help improve contextual understanding and robustness.

Current LLMs are insufficient for Chinese-specific fact-checking tasks, especially those requiring precision or cultural expertise. Our research shows that even Chinese-focused LLMs struggle with certain culturally specific issues, such as lunar calendar calculations. In this area, LLMs achieved an accuracy rate of only 19.6%, highlighting their difficulty in processing culturally nuanced knowledge. Potential improvements could include culturally specific data and domain-specific fine-tuning.

7 Conclusion

We investigate LLMs' deficiencies in fact-checking real-world Chinese misinformation by: 1) proposing CANDY, the first benchmark tailored for this task; 2) introducing a fine-grained taxonomy and the large-scale CANDYSET dataset to evaluate LLMs' performance across real-time and outdated contexts; and 3) assessing sixteen mainstream LLMs to uncover key challenges and limitations in their fact-checking capabilities. Our work serves as a valuable resource, offering insights and guidance for future advancements in this field.

Limitations

Sensitivity of Prompts. Similar to other studies on prompting large language models (LLMs) (Deng et al., 2023; Zhang et al., 2024), the evaluation results are likely to be sensitive to the prompts used. Although we utilize four distinct prompts and present the average outcomes, it is difficult to claim that these are the most optimal for our particular task. In fact, fine-tuning prompts for this specific application remains a substantial challenge and an important direction for future research.

Limited LLMs for annotation. Unlike the fact-checking conclusion task, which experiments with 16 LLMs (8 open-source and 8 closed-source) on the entire CANDYSET dataset, the fact-checking explanation task was limited by the cost of manual analysis and labeling. As a result, only 9 models (6 closed-source and 3 open-source) were selected to generate analysis and labels on a randomly chosen subset of 4.5k data entries. If more labeling resources become available in the future, we plan to extend this analysis to the remaining models.

Ethics Statement

Our work introduces the CANDYSET dataset, which contains real-world Chinese misinformation. We acknowledge the ethical implications of handling and disseminating misinformation, and we are committed to ensuring that our research is conducted responsibly and ethically. The primary goal of this research is to evaluate and improve the performance of LLMs in identifying and mitigating the impact of misinformation. By testing LLMs on this dataset, we aim to advance the understanding of how these models can be refined to better discern factual accuracy and provide reliable information. Therefore, we emphasize that this dataset should only be used within the scope of research aimed at combating misinformation, and not for spreading or endorsing false information. We advise researchers and practitioners to employ this dataset responsibly, ensuring that the findings contribute positively to the development of more robust and truthful LLMs. We are committed to transparency in our methodologies and findings, and we welcome feedback from the community to improve our approaches. In all studies involving human subjects, we diligently followed IRB approval protocols. Each human annotator received a compensation of $200 for their contributions.

References

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

BaiChuan. 2024. baichuan4-turbo.

Mars Gokturk Buchholz. 2023. Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint arXiv:2306.08190*.

Ju Cao, Jiafeng Guo, Xueqi Li, Zitao Jin, Han Guo, and Jiaming Li. 2018. Automatic rumor detection on microblogs: A survey. *arXiv preprint arXiv:1807.03505*.

Recep Firat Cekinel and Pinar Karagoz. 2024. Explaining veracity predictions with evidence summarization: A multi-task model approach. *arXiv preprint arXiv:2402.06443*.

Qiang Cheng, Tianwei Sun, Wenhao Zhang, Sheng Wang, Xiaotang Liu, Min Zhang, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Jeremy Cole, Michael Zhang, Daniel Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543, Singapore. Association for Computational Linguistics.

Anik Das, Hui Liu, Veselin Kovatchev, and Matthew Lease. 2023. The state of human-centered nlp technology for fact-checking. *Information Processing & Management*, 60(2):103219.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022a. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022b. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Z. Dong, T. Tang, J. Li, W. X. Zhao, and J. R. Wen. 2024. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2086–2099.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

R. Friel and A. Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *Unpublished Manuscript or Preprint (if applicable)*. Accessed on [date if applicable].

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503.

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023. Leveraging chat-gpt for efficient fact-checking. Available at: `https://doi.org/10.31234/osf.io/qnjkf`.

Y. L. Hsu, J. N. Chen, Y. F. Chiang, S. C. Liu, A. Xiong, and L. W. Ku. 2024. Enhancing perception: Refining explanations of news claims with llm conversations. In *Findings of the Association for Computational Linguistics: NAACL*, pages 2129–2147.

B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, and P. Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, and T. Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Y. Huang, K. Shu, P. S. Yu, and L. Sun. 2024. From creation to clarification: Chatgpt's journey through the fake news quagmire. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 513–516.

S. Kang, G. An, and S. Yoo. 2024. A quantitative and qualitative evaluation of llm-based explainable fault localization. *Proceedings of the ACM on Software Engineering*, 1(FSE):1424–1446.

Wei-Yu Kao and An-Zi Yen. 2024. How we refute claims: Automatic fact-checking through flaw identification and explanation. *arXiv preprint arXiv:2401.15312*.

K Kim, S Lee, K H Huang, et al. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

J. Li, J. Chen, R. Ren, X. Cheng, W. X. Zhao, J. Y. Nie, and J. R. Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yifan Li and ChengXiang Zhai. 2023. An exploration of large language models for verification of news headlines. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, page 197206.

X. Liang, S. Song, S. Niu, Z. Li, F. Xiong, B. Tang, and H. Deng. 2023. Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. *arXiv preprint arXiv:2311.15296*.

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

Dor Muhlgay, Ori Ram, Inbal Magar, et al. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66. Association for Computational Linguistics.

OpenAI. 2024a. gpt-3.5.

OpenAI. 2024b. Hello gpt-4o.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics.

R. Patil and V. Gudivada. 2024. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074.

E. Porter, R. B. Scott, T. J. Wood, and R. Zhandayeva. 2024. Correcting misinformation about the russia-ukraine war reduces false beliefs but does not change views about the war. *PLOS ONE*, 19(9):e0307090.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

M. Sundriyal, T. Chakraborty, and P. Nakov. 2023. From chaos to clarity: Claim normalization to empower fact-checking. *arXiv preprint arXiv:2310.14338*.

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*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sander van der Linden. 2022. Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28(3):460–467.

T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, and T. Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

I. Vykopal, M. Pikuliak, S. Ostermann, and M. Šimko. 2024. Generative large language models in automated fact-checking: A survey. *arXiv preprint arXiv:2407.02351*.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, and Arman Cohan. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7534–7550.

H. Wan, S. Feng, Z. Tan, H. Wang, Y. Tsvetkov, and M. Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Binjie Wang, Ethan Chern, and Pengfei Liu. 2023. Chinesefacteval: A factuality benchmark for chinese llms. Technical report, GAIR-NLP.

Ning Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. 2014. A comprehensive survey to face hallucination. *International Journal of Computer Vision*, 106:9–30.

Y. Wang, Q. Liu, and C. Jin. 2024. Is rlhf more difficult than standard rl? a theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 36.

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.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *ArXiv preprint*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, and T. S. Chua. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. CLAMBER: A benchmark of identifying and clarifying ambiguous information needs in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10746–10766, Bangkok, Thailand. Association for Computational Linguistics.

A Taxonomy Table

Dimension	Category	Definition	Example
Faithfulness Hallucination	Instruction Inconsistency	Refers to the LLM's output that deviate from user's directive.	Input: Beijing time does not come from Beijing. Output: (Invalid). A news report states a new type of plant can grow in saltwater...
	Logical Inconsistency	Refers to instances where LLMs explanation contains internal logical conflicts.	Input: YaYa's stay in the United States lasts for 15 years. Output: Nonrumor. Yaya stay in USA from 2003 to 2023...totaly 15 years.
	Context Inconsistency	Refers to the LLM's output being inconsistent with user-provided context.	Input: Drinking a large amount of water after exercise is not recommended. Output: Rumor. Completely prohibiting drinking water after exercise is incorrect...
Factuality Hallucination	Factual Fabrication	Refers to the LLM's output that fabricates rationales for analysis without relying on real-world information.	Input: A major accident occurred on the Jingkun Expressway at the Huo County section (Fabricatied misinformation). Output: Nonrumor. According to Chinese media reports, this traffic accident indeed occurred...
	Factual Inconsistency	Refers to the LLM's output contains facts that can be grounded in real-world information, but present contradictions.	Input: Houttuynia cordata contains Aristolochia acid II, which is toxic and carcinogenic (Actually acid I). Output: Nonrumor. IARC has classified Aristolochia acid II as a human carcinogen...
Reasoning Inadequacy	Overgeneralized Reasoning	Refers to the tendency of a LLM to produce speculative rationales based on overly broad or superficial criteria.	Input: Over 100 billion yuan in agricultural subsidies will support 23 eligible projects. Output: Nonrumor. A figure exceeding 100 billion is not exaggerated...
	Under Informativeness	Refers to the tendency of a LLM to exhibit excessive rigor or restraint, failing to provide more contextually valuable content.	Input: Moderate playing of Mahjong can prevent senile dementia. Output: Rumor. The WHO have not announced that...

Table 6: Taxonomy.

B Details of Dataset Construction

This section primarily describes the details of the data gathering pipeline we proposed as figure 5.

B.1 Data Crawling

Initially, we crawled data from authoritative Chinese fact-checking agencies. We mainly explain the process of extracting data from the China Internet Joint Rumor Refuting Platform. This platform is a active Chinese fact-checking website listed by Duke Reporters[5] . This platform not only provides timely refuting of misinformation that have recently (usually the day before) attracted attention on the internet, but also features a "*Daily Popular Science*" segment, which can serve as a source of genuine claims to ensure the dataset's balance. Specifically, we collected the following information, including authentic or deceptive items and their corresponding facts and timestamps, covering content from January 2023 to October 2024. These claims cover multiple domains, including politics, health, science, society, life, culture and disasters.

B.2 Data Normalization

Data normalization encompasses data cleaning and normalization (Sundriyal et al., 2023). Initially, we manually inspect and remove low-quality data, such as those with insufficient background information and unverifiable subjective rumors (Cao et al., 2018). Given that the crawled data includes well-reasoned truths from authoritative sources, we summarize these truths as fact-checking gold evidence related to claim verification and label the corresponding claims (Hanselowski et al., 2019). In this process, we use GPT-4o for initial data preprocessing and labeling, followed by manual verification.

B.3 Data Augmentation

To assess LLMs robustness and enhance label balance, we introduced data augmentation techniques like subtle modifications to existing claims, to observe changes in responses. These modifications involved altering event details or adjusting the veracity of statements using negations. For instance, when we replaced the entity in the statement "*The 'Food Safety National Standard - Contaminants in Food' stipulates that the limit for pickled vegetables is 20 milligrams per kilogram*" with "*toona sinensis*", the model was unable to accurately identify the change, leading to an occurrence of Faithfulness Hallucination.

[5]www.reporterslab.org/fact-checking/

B.4 Data Validation

To ensure a high-quality dataset, we carefully performed manual validation on both the labels and the gold evidence. Firstly, to validate the ground truth labels, we performed a sampling check by randomly selecting 3% of the dataset—approximately 600 entries—for detailed review. Each entry was re-annotated by five independent annotators to assess consistency and accuracy. To quantify inter-annotator agreement, we calculated the Fleiss' Kappa score (Fleiss, 1971), which yielded a value of 0.75. This indicates substantial agreement, confirming the reliability of the annotations. Additionally, we evaluated whether the gold evidence provided with each claim was sufficient to accurately support or refute the claim. A separate group of annotators reviewed these sampled entries to verify that the evidence was comprehensive and relevant. This dual-layered approach not only checked for annotation consistency but also assessed the informativeness and adequacy of the evidence. Through this process, we maintained a high standard of data quality, ensuring that the dataset is reliable for use in real-world fact-checking applications.

B.5 Fact-checking Explanation annotation

To perform a more in-depth analysis of the accuracy of LLM-generated explanations in the fact-checking task, we recruited three experienced annotators with master's degrees in computer science and technology, proficiency in English, and extensive experience in data annotation. These annotators were responsible for classifying errors in the LLM-generated explanations according to our taxonomy, which is detailed in Table 6. The results of the annotation process are visualized in Figure 6. To ensure the reliability and quality of the annotations, each explanation was independently labeled by two annotators. In instances where there were significant discrepancies between their annotations, a third annotator was consulted to review the explanations and resolve the differences through discussion. This additional layer of review helped mitigate bias and ensured that the final annotations were as accurate as possible. To quantify the consistency of the annotations, we calculated the Fleiss' Kappa score, which measures inter-annotator agreement. The resulting score of 0.76 indicates a substantial level of agreement, suggesting that the an-

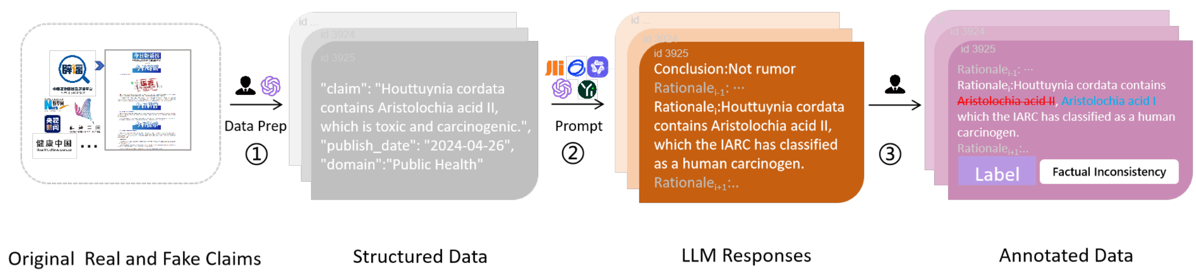

Figure 5: Data gathering pipeline. our data gathering pipeline includes 3 steps: 1) Data collection and pre-processing. 2) Response generation. 3) Human annotation.

Platform	English Name	Link	Count
中国互联网联合辟谣平台	China Internet United Rumor Refutation Platform	https://www.piyao.org.cn/	2172
新华社	Xinhua News Agency	https://www.xinhuanet.com/	1255
科普中国	Science Popularization China	https://www.kepuchina.cn/	595
央视新闻	CCTV News	https://news.cctv.com/	497
人民网科普	People's Daily Online Science Popularization	https://kpzg.people.com.cn/	465
健康中国	Healthy China	https://www.nhc.gov.cn/	255
科学辟谣	Science Rumor Refutation	https://www.kepuchina.cn/	210
上海网络辟谣	Shanghai Network Rumor Refutation	https://piyao.jfdaily.com/	168
中国新闻网	China News Service	https://www.chinanews.com.cn/	144
网信中国	Cyberspace Administration of China	https://www.cac.gov.cn/	131

Table 7: Top 10 Sources of CANDYSET

notation process was both reliable and robust. This high level of agreement provides confidence in the validity of the annotated data and supports the subsequent analysis of LLM-generated explanations in the context of fact-checking.

C Implementation Details

We conduct all our experiments using a single Nvidia RTX A100 GPU for the 6 and 7B size LLMs, two A100 GPUs for the 9B and 13B size LLMs, and four A100 GPUs for the 70B and 72B size LLMs. For these open-source LLMs, we utilize the Xinference framework. For all LLMs, we employ nucleus sampling with a temperature of 0.7 and a top-p value of 0.95, allowing for a maximum of 10 iterations per stage with human programmers. For the accuracy and F1 metrics, we calculate it using the micro average method.

D Additional Results

We present additional results here.

E Chinese Figures

We present the English version of the main text images here.

F LLMs Employed in This Research

The large language models (LLMs) employed in this research and their respective knowledge cut-off dates and access links are shown in the table below.

G Prompt Design

Following Deng et al. (2023), we propose four prompting schemes for the fact-checking conclusion task:

1. **Zero-shot w/o CoT**, where LLMs are prompted to directly draw conclusions;
2. **Zero-shot w/ CoT** (Wei et al., 2022), where LLMs first perform a factual analysis, explaining their reasoning before making a conclusion;
3. **Few-shot w/o CoT** (Dong et al., 2022a), where LLMs are given a few examples to guide their conclusions;

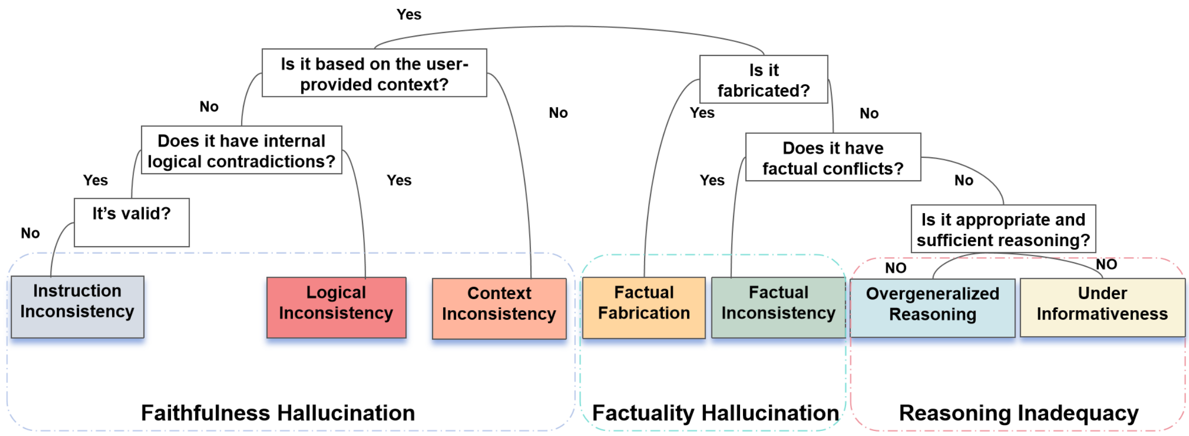

Figure 6: Decision Tree for Annotation

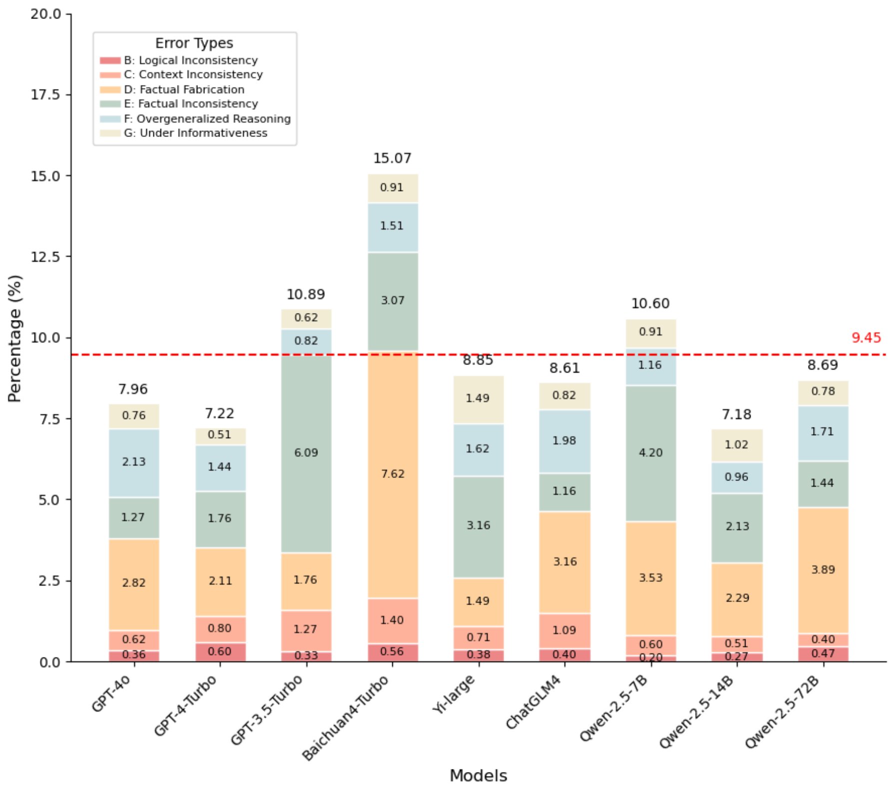

Figure 7: Overall Error density of models.

核查如下主张是否符合事实：金字塔是外星人修建的。

Factual Fabrication	金字塔是外星人储存能量的装置。
Factual INC.	金字塔不是外星人修建的，其中最大的是哈夫拉金字塔 (胡夫金字塔)。
Instruction INC.	吃转基因食品会致癌。(曲解任务为生成不符事实信息。)
Logical INC.	金字塔不是外星人修建的，其大约建于公元前2580年至前2560年期间，距今约2500年 (4500年)。
Context INC.	金字塔是由花岗岩和石灰石建造而成的。(曲解待核实主张为金字塔是用什么修建的。)
Overgeneralized Reasoning	金字塔的建造是外星人完成的，因为金字塔的大小和重量远远超出了人类当时的技术能力。(依据过于表面且不可靠。)
Under Informativeness	没有权威证明金字塔是否由外星人修建。(未提供充分信息。)

Figure 8: Demonstrating our taxonomy with cases in Chinese. (In English: Fig. 1)

Declarative Claim	吃转基因食品会致癌。[虚假信息] 非谣言。根据世界卫生组织（WHO）和其他权威机构的说法，转基因食品确实被证明会导致癌症。[Factual Fabrication]
Counter Claim	吃转基因食品不会导致癌症。 非谣言。目前没有科学共识支持这一说法。[Stance change]
Interrogative Form	吃转基因食品会致癌吗? 谣言。转基因食品的测试涵盖毒性、过敏原性和营养等方面。权威机构的研究表明，食用转基因食品不会导致癌症。[Honest & Informative]

Figure 9: The influence of claim framing strategies on fact-checking outputs. (In English: Fig. 4)

Model Name	**Cut-off Date**	**Link**
GPT-4o	2023-11	GPT-4o
GPT-4-Turbo	2023-5	GPT-4-Turbo
GPT-3.5-Turbo	2021-10	GPT-3.5-Turbo
Gemini-1.5-pro	2023-11	Gemini-1.5-pro
Baichuan4-Turbo	2024-04	Baichuan4-Turbo
ChatGLM4	2022-10	ChatGLM4
Yi-large	2023-6	Yi-large
Deepseek-v3	2024-7	Deepseek-v3
Yi-1.5-6B	2024-5	Yi-1.5-6B
Qwen-2.5-7B	2023-10	Qwen-2.5-7B
Llama-3.2-7B	2023-12	Llama-3.2-7B
GLM4-9B	2023-10	GLML4-9B
Yi-1.5-9B	2024-5	Yi-1.5-9B
Qwen-2.5-14B	2023-10	Qwen-2.5-14B
Llama-3.2-72B	2023-12	Llama-3.2-72B
Qwen-2.5-72B	2023-10	Qwen-2.5-72B

Table 8: LLMs Overview

4. **Few-shot w/ CoT** (Dong et al., 2022b), where LLMs, after analyzing examples of misinformation, provide conclusions along with explanations.

For the fact-checking explanation task, we use a few-shot CoT approach, prompting the LLM to generate five distinct explanatory rationales for the conclusion.

In the few-shot setting, we provide two randomly selected examples: one authentic and the other misinformation. We carefully selected three different prompts and evaluated all LLMs using these prompts. To ensure the statistical significance of the findings, we present the average performance across all prompts.

Our analysis of Chinese fact-checking conclusions and explanations aims to probe the knowledge embedded within the LLM, resembling a closed-book QA process with questions from the `CANDYSET` (Roberts et al., 2020). The prompt design is crucial in this context. We adopt a **Few-shot with Chain of Thought (CoT)** (Dong et al., 2022b) approach, where LLMs are evaluated by providing examples along with their corresponding fact-checking responses. For each claim, two randomly selected examples—one representing misinformation and the other genuine news—are provided. Following the methodology outlined by Li et al. (2024), we instruct LLMs to generate five mutually exclusive rationales, incorporating factual statements for detailed, sentence-level error analysis.

Additionally, techniques such as **role specification** (Li and Zhai, 2023) have been shown to improve performance. In our study, we assign LLMs the specific role of rigorous fact-checkers, ensuring they are tasked with the responsibility of meticulously analyzing the information's veracity. Through carefully constructed prompts, we enhance the transparency of the LLM's reasoning process, allowing for a clearer understanding of how decisions about the truthfulness of information are made.

Zero-shot and w/o CoT Prompt for Fact-Checking Conclusion (Chinese)

你是一位极其严格的事实核查专家，你会收到用户输入的事件信息，其中日期信息可能略晚于claim实际发布日期，你需要直接给出结论。其中结论只能是：谣言/非谣言。
输出格式如下：
结论：
现在，我将提供一个新的事件信息，请你根据以上格式给出结论和分析。
事件信息：
"claim": "claim",
"publish_date": "publish_date",

Zero-shot and w/o CoT Prompt for Fact-Checking Conclusion (English)

You are an extremely strict fact-checking expert. You will receive event information from users, where the date provided may be slightly later than the actual publication date of the claim. You need to provide the conclusion directly, which can only be: rumor or non-rumor.

Output Format:
Conclusion:

Now, I will provide a new event information. Please give a conclusion and analysis based on the above format.
Event Information:
"claim": "claim",
"publish_date": "publish_date"

Few-shot and w/o CoT Prompt for Fact-Checking Conclusion (Chinese)

你是一位极其严格的事实核查专家，你会收到用户输入的事件信息，其中日期信息可能略晚于claim实际发布日期，你需要直接给出结论。其中结论只能是：谣言/非谣言。
示例如下：
用户输入: "claim": "吃竹炭食物能排毒养颜。", "publish_date": "2019-10-08"
回复：
结论：谣言
用户输入: "claim": "没签劳动合同的职工受伤后可以申请工伤认定。", "publish_date": "2023-12-5"
回复：
结论：非谣言

输出格式如下：
结论：
现在，我将提供一个新的事件信息，请你根据以上格式给出结论和分析。
事件信息：
"claim": "claim",
"publish_date": "publish_date",

Few-shot and w/o CoT Prompt for Fact-Checking Conclusion (English)

You are an extremely strict fact-checking expert. You will receive event information from users, where the date provided may be slightly later than the actual publication date of the claim. You need to provide the conclusion directly, which can only be: rumor or non-rumor.

Examples are as follows:
User input: "claim": "Eating bamboo charcoal foods can detoxify and improve skin appearance.", "publish_date": "2019-10-08"
Response:
Conclusion: Rumor
User input: "claim": "Employees who have not signed a labor contract can still apply for work injury recognition after being injured.", "publish_date": "2023-12-5"
Response:
Conclusion: Non-rumor

Output format:
Conclusion:
Now, I will provide a new event information. Please give the conclusion and analysis according to the above format.
Event information:
"claim": "claim",
"publish_date": "publish_date",

Zero-shot and w CoT Prompt for Fact-Checking Conclusion (Chinese)

你是一位极其严格的事实核查专家，你会收到用户输入的事件信息，其中日期信息可能略晚于claim实际发布日期，你需要先给出五个事实性陈述作为理由，再给出结论。其中结论只能是：谣言/非谣言。
输出格式如下：
理由：
结论：
现在，我将提供一个新的事件信息，请你根据以上格式给出结论和分析。
事件信息：
"claim": "claim",
"publish_date": "publish_date",

Zero-shot and w CoT Prompt for Fact-Checking Conclusion (English)

You are an extremely strict fact-checking expert. You will receive event information from users, where the date provided may be slightly later than the actual publication date of the claim. You need to first provide five factual statements as reasons, and then give a conclusion. The conclusion can only be: rumor or non-rumor.
Output Format:
Rationales:
Conclusion:

Now, I will provide a new event information. Please give a conclusion and analysis based on the above format.
Event Information:
"claim": "claim",
"publish_date": "publish_date"

Few-shot and w CoT Prompt for Fact-Checking Conclusion (Chinese)

你是一位极其严格的事实核查专家，你会收到用户输入的事件信息，其中日期信息可能略晚于claim实际发布日期，你需要首先生成五个事实性陈述作为理由，然后根据理由给出结论。其中结论只能是：谣言/非谣言。
示例如下：
用户输入: "claim": "吃竹炭食物能排毒养颜。", "publish_date": "2019-10-08"
回复：
理由：
1. 竹炭被广泛用于食品和医疗，但其排毒效果缺乏科学证据支持。
2. 根据多项研究，人类的肝脏和肾脏自然具有排毒功能，额外摄入竹炭并未得到临床验证的排毒效果。
3. 竹炭食品可能会影响营养物质的吸收，例如维生素和矿物质，也可能带来健康风险。
4. 一些国家和地区对竹炭的食品使用有法律限制，需遵循相关法规。
5. 目前没有权威机构发布证实竹炭食品对美容或健康有益的声明。
用户输入: "claim": "没签劳动合同的职工受伤后可以申请工伤认定。", "publish_date": "2023-12-5"
结论：谣言

回复：
理由：
1.根据《工伤保险条例》，职工与用人单位建立劳动关系后，即使未签订书面劳动合同，在工作时间和工作场所内因工作原因受到事故伤害的，应认定为工伤。
2.未签订劳动合同的职工在发生工伤时，可通过提供其他证明材料（如工资支付凭证、工作证等）来证明存在劳动关系，进而申请工伤认定。
3.劳动和社会保障部门负责工伤认定工作，会根据实际情况进行调查核实，确认是否存在劳动关系以及是否符合工伤认定条件。
4.用人单位未与职工签订劳动合同属于违法行为，职工有权向劳动监察部门投诉，要求用人单位补签劳动合同或赔偿相应损失。
5. 工伤认定不仅涉及劳动者权益保护，也是企业社会责任的重要体现，有助于维护社会稳定和谐。
结论：非谣言

输出格式如下：
理由：
结论：
现在，我将提供一个新的事件信息，请你根据以上格式给出结论和分析。
事件信息：
"claim": "claim",
"publish_date": "publish_date",

Few-shot and w CoT Prompt for Fact-Checking Conclusion (English)

You are an extremely strict fact-checking expert. You will receive event information from users, where the date provided may be slightly later than the actual publication date of the claim. You need to first generate five factual statements as reasons and then draw a conclusion based on those reasons. The conclusion can only be: rumor or non-rumor.

Example:
User Input:
"claim": "Eating bamboo charcoal food can detoxify and beautify.", "publish_date": "2019-10-08"
Rationales:
1.Bamboo charcoal is widely used in food and medicine, but its detoxification effects lack scientific evidence.
2.According to multiple studies, the human liver and kidneys naturally have detoxification functions, and additional intake of bamboo charcoal has not been clinically validated for detoxification effects.
3.Bamboo charcoal food might affect the absorption of nutrients such as vitamins and minerals, and could pose health risks.
4.Some countries and regions have legal restrictions on the use of bamboo charcoal in food, which must be followed.
5.No authoritative institution has released statements confirming that bamboo charcoal food is beneficial for beauty or health.
Response:
Conclusion: Rumor

User Input:
"claim": "Workers injured without a signed labor contract can apply for work injury recognition.", "publish_date": "2023-12-5"
Rationales:
1.According to the "Work Injury Insurance Regulations," if a worker has established a labor relationship with the employer, even without a signed written contract, they should be recognized as having a work injury if injured in the workplace due to work-related Rationales.
2.Workers without a signed labor contract can provide other evidence (such as wage payment receipts, work IDs, etc.) to prove the existence of a labor relationship and apply for work injury recognition.
3.The labor and social security department is responsible for work injury recognition and will investigate and verify whether there is a labor relationship and if it meets the conditions for work injury recognition.
4.Employers failing to sign labor contracts with workers is illegal, and workers have the right to file complaints with labor inspection departments, requesting the employer to sign a contract or compensate for relevant losses.
5.Work injury recognition not only involves protecting worker rights but also reflects corporate social responsibility, helping maintain social stability and harmony.
Response:
Conclusion: Non Rumor

Output Format:
Rationales:
Conclusion:

Now, I will provide a new event information. Please give a conclusion and analysis based on the above format.
Event Information:
"claim": "claim",
"publish_date": "publish_date"

Prompt for Fact-Checking Explanation (Chinese)

你是一位极其严格的事实核查专家，你会收到用户输入的事件信息，其中日期信息可能略晚于claim实际发布日期，你需要首先生成五个事实性陈述作为理由，然后根据理由给出结论。其中结论只能是：谣言/非谣言。
示例如下：
用户输入: "claim": "吃竹炭食物能排毒养颜。", "publish_date": "2019-10-08"
回复：
理由：
1. 竹炭被广泛用于食品和医疗，但其排毒效果缺乏科学证据支持。
2. 根据多项研究，人类的肝脏和肾脏自然具有排毒功能，额外摄入竹炭并未得到临床验证的排毒效果。
3. 竹炭食品可能会影响营养物质的吸收，例如维生素和矿物质，也可能带来健康风险。
4. 一些国家和地区对竹炭的食品使用有法律限制，需遵循相关法规。
5. 目前没有权威机构发布证实竹炭食品对美容或健康有益的声明。
用户输入: "claim": "没签劳动合同的职工受伤后可以申请工伤认定。", "publish_date": "2023-12-5"
结论：谣言

回复：
理由：
1.根据《工伤保险条例》，职工与用人单位建立劳动关系后，即使未签订书面劳动合同，在工作时间和工作场所内因工作原因受到事故伤害的，应认定为工伤。
2.未签订劳动合同的职工在发生工伤时，可通过提供其他证明材料（如工资支付凭证、工作证等）来证明存在劳动关系，进而申请工伤认定。
3.劳动和社会保障部门负责工伤认定工作，会根据实际情况进行调查核实，确认是否存在劳动关系以及是否符合工伤认定条件。
4.用人单位未与职工签订劳动合同属于违法行为，职工有权向劳动监察部门投诉，要求用人单位补签劳动合同或赔偿相应损失。
5. 工伤认定不仅涉及劳动者权益保护，也是企业社会责任的重要体现，有助于维护社会稳定和谐。
结论：非谣言

输出格式如下：
理由：
结论：
现在，我将提供一个新的事件信息，请你根据以上格式给出结论和分析。
事件信息：
"claim": "claim",
"publish_date": "publish_date",

1086

Prompt for Fact-Checking Explanation (English)

You are an extremely strict fact-checking expert. You will receive event information from users, where the date provided may be slightly later than the actual publication date of the claim. You need to first generate five factual statements as reasons and then draw a conclusion based on those reasons. The conclusion can only be: rumor or non-rumor.

Example:
User Input:
"claim": "Eating bamboo charcoal food can detoxify and beautify.", "publish_date": "2019-10-08"
Rationales:
1.Bamboo charcoal is widely used in food and medicine, but its detoxification effects lack scientific evidence.
2.According to multiple studies, the human liver and kidneys naturally have detoxification functions, and additional intake of bamboo charcoal has not been clinically validated for detoxification effects.
3.Bamboo charcoal food might affect the absorption of nutrients such as vitamins and minerals, and could pose health risks.
4.Some countries and regions have legal restrictions on the use of bamboo charcoal in food, which must be followed.
5.No authoritative institution has released statements confirming that bamboo charcoal food is beneficial for beauty or health.
Response:
Conclusion: Rumor

User Input:
"claim": "Workers injured without a signed labor contract can apply for work injury recognition.", "publish_date": "2023-12-5"
Rationales:
1.According to the "Work Injury Insurance Regulations," if a worker has established a labor relationship with the employer, even without a signed written contract, they should be recognized as having a work injury if injured in the workplace due to work-related Rationales.
2.Workers without a signed labor contract can provide other evidence (such as wage payment receipts, work IDs, etc.) to prove the existence of a labor relationship and apply for work injury recognition.
3.The labor and social security department is responsible for work injury recognition and will investigate and verify whether there is a labor relationship and if it meets the conditions for work injury recognition.
4.Employers failing to sign labor contracts with workers is illegal, and workers have the right to file complaints with labor inspection departments, requesting the employer to sign a contract or compensate for relevant losses.
5.Work injury recognition not only involves protecting worker rights but also reflects corporate social responsibility, helping maintain social stability and harmony.
Response:
Conclusion: Non Rumor

Output Format:
Rationales:
Conclusion:

Now, I will provide a new event information. Please give a conclusion and analysis based on the above format.
Event Information:
"claim": "claim",
"publish_date": "publish_date"

H Annotation Examples

English Example of Faithfulness Hallucination

id: 882,
claim: In cases of bacterial infection, antibiotics can be used to treat COVID-19 patients.,
publish_date: 2023/1/3,
label: 0,
source: China Media Group "China's Voice" Official WeChat Account,
gold evidence: Zhang Bo stated, "In clinical practice, we find that some patients have co-existing bacterial infections, or secondary bacterial infections caused by COVID-19. In such cases, we can choose antibiotics or antimicrobial drugs.",
domain: Public Health,
model: GPT-4o,
analysis: [
1. Antibiotics are specifically used to treat bacterial infections, while COVID-19 is caused by a coronavirus, making antibiotics ineffective against viral infections.,
2. Both the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) have explicitly stated that antibiotics should not be used to prevent or treat COVID-19.,
3. Overuse of antibiotics can lead to increased antibiotic resistance, which poses a serious threat to public health.,
4. Treating COVID-19 requires the use of antiviral medications and supportive care based on the specific condition, rather than antibiotics.,
5. Only in cases where bacterial complications occur might doctors consider the use of antibiotics, based on the specific situation.],
label_p: 1,
Error Type: C:Context Inconsistency

id: 462,
claim: "Ya Ya's" stay in the United States was for 15 years.,
publish_date: 2023/4/24,
label: 1,
source: null,
gold evidence: "The agreement for 'Ya Ya's' stay in the United States was initially set for 10 years and was extended by another 10 years after expiring in 2013.",
domain: Society,
model: Qwen-2.5-72B,
analysis: [
1. "Ya Ya" is a giant panda born in China. In April 2003, "Ya Ya" and another panda, "Le Le," were sent to the Memphis Zoo in Tennessee, USA, under a 10-year loan agreement.,
2. According to the agreement between China and the United States, after the initial 10-year term expired in 2013, the loan agreement for "Ya Ya" and "Le Le" was renewed for another 10 years, extending their stay until April 7, 2023, making the total stay 15 years.,
3. This agreement's details were officially confirmed by the Chinese Association of Wildlife Conservation and the Memphis Zoo.,
4. During "Ya Ya" and "Le Le's" stay in the United States, both countries collaborated on panda conservation and research, showcasing mutual efforts in wildlife protection.,
5. In early 2023, "Le Le" sadly passed away, raising concerns about "Ya Ya's" health, which eventually led to the decision to bring "Ya Ya" back to China earlier than planned. However, this did not alter the originally agreed 15-year term.],
label_p: 0,
Error Type: B:Logical Inconsistency

1089

Chinese Example of Faithfulness Hallucination

id: 882,
claim: 在细菌感染的情况下，可以用抗生素治疗新冠患者。,
publish_date: 2023/1/3,
label: 0,
source: 中央广电总台中国之声微信公众号,
gold evidence: 张波称，“在临床中我们会发现，有些患者合并细菌感染，或者说新冠引起的继发细菌感染，在这种情况下，我们是可以选择抗生素或抗菌药物的。”,
domain: 公卫,
model: GPT-4o,
analysis: [
1. 抗生素专门用于治疗细菌感染，而COVID-19是由冠状病毒引起的，抗生素对病毒感染无效。,
2. 世界卫生组织（WHO）和美国疾病控制与预防中心（CDC）都明确指出，抗生素不能用于预防或治疗COVID-19。,
3. 过度使用抗生素可能导致抗生素耐药性增加，这对公共健康构成威胁。,
4. 治疗COVID-19需根据具体病情使用抗病毒药物和支持性治疗，而非抗生素。,
5. 只有在出现细菌感染并发症时，医生才可能根据具体情况考虑使用抗生素治疗。],
label_p: 1,
Error Type: C:Context Inconsistency

id: 462,
claim: “丫丫”在美国的旅居期限为15年。,
publish_date: 2023/4/24,
label: 1,
source: null,
gold evidence: “丫丫”旅居美国协议期限为10年，2013年到期后又延长了10年。,
domain: 社会,
model: Qwen-2.5-72B,
analysis: [
1. “丫丫”是一只在中国出生的大熊猫，于2003年4月与另一只大熊猫“乐乐”一起被送往美国田纳西州孟菲斯动物园，开始了为期10年的租借期。,
2. 根据中美双方协议，2013年租借期满后，两国再次签署了延长协议，将“丫丫”和“乐乐”的旅美期限延长至2023年4月7日，总计15年。,
3. 这一协议的内容得到了中国野生动物保护协会和美国孟菲斯动物园的官方确认。,
4. 在“丫丫”和“乐乐”的旅美期间，两国在大熊猫保护与科研合作方面进行了多次交流与合作，体现了双方在野生动物保护领域的共同努力。,
5. 2023年初，“乐乐”不幸去世，引发了关于“丫丫”健康状况的关注，最终促成了“丫丫”提前返回中国的决定，但这并不改变原定的15年旅居期限的事实。],
label_p: 0,
Error Type: B:Logical Inconsistency

English Example of Factuality Hallucination

id: 199,
claim: A criminal case occurred in Jilin Province, where a butcher killed 12 people due to a land dispute.,
publish_date: 2024/8/14,
label: 1,
source: China Internet Joint Rumor Debunking Platform,
gold evidence: After investigation by the police, it was found that no such case occurred at the location, and this information was entirely fabricated by internet users to gain attention and attract traffic.,
domain: Society,
model: Baichuan4-Turbo,
analysis: [
1. A criminal case indeed occurred in Jilin Province, and it has drawn widespread local attention.,
2. According to the police report, the suspect developed resentment due to a land dispute, leading to this tragedy.,
3. The number of victims was 12, a figure confirmed by the authorities.,
4. The suspect was a butcher, a detail also confirmed by the authorities.,
5. The case is under further investigation, and the responsible parties will be punished according to the law.],
label_p: 0,
Error Type: D:Factual Fabrication

id: 1394,
claim: Lunar volcanic activity stopped 3 billion years ago.,
publish_date: 2023/1/20,
label: 1,
source: Science China (Popular Science Platform),
gold evidence: It was discovered that lunar volcanic activity actually persisted until 2.03 billion years ago, rather than ceasing 3 billion years ago as previously believed.,
domain: Science,
model: Baichuan4-Turbo,
analysis: [
1. Scientists have found through the analysis of lunar rock samples brought back by the Apollo missions that volcanic activity on the moon indeed peaked around 3 billion years ago and then gradually declined.,
2. So far, there is no conclusive evidence suggesting that the moon had active volcanic activity after 3 billion years ago.,
3. Observations and explorations of the lunar surface, including the density and distribution of impact craters, support the view that lunar volcanic activity ceased billions of years ago.,
4. Compared to Earth, the moon's geological activity is very weak, and its rapid internal heat loss led to an earlier end to volcanic activity.,
5. Research by NASA and other international space agencies supports the presence of ancient lunar volcanic activity, which had essentially ended by 3 billion years ago.],
label_p: 0,
Error Type: E:Factual Inconsistency

Chinese Example of Factuality Hallucination

id: 199,
claim: 吉林省发生了一起刑事案件，一名屠夫因为土地纠纷杀害了12人。,
publish_date: 2024/8/14,
label: 1,
source: 中国互联网联合辟谣平台,
gold evidence: 经警方调查发现，案发地并未发生此类案件，这则消息纯属子虚乌有，系网民为了博取关注、吸引流量而编造发布的谣言。,
domain: 社会,
model: Baichuan4-Turbo,
analysis: [
1. 吉林省确实发生了这起刑事案件,该事件在当地引起了广泛关注。,
2. 据警方通报，犯罪嫌疑人因土地纠纷产生不满情绪，最终导致了悲剧的发生。,
3. 被害人数为12人，这一数字得到了官方确认。,
4. 犯罪嫌疑人为一名屠夫，这一点也得到了官方证实。,
5. 此案件正在进一步调查中，相关责任人将依法受到严惩。],
label_p: 0,
Error Type: D:Factual Fabrication

id: 1394,
claim: 月球的火山活动在 30 亿年前就停止了。,
publish_date: 2023/1/20,
label: 1,
source: 科普中国,
gold evidence: 发现月球的火山活动其实一直持续到了 20.3亿年前，并不是之前认为的 30 亿年前就死了。,
domain: 科学,
model: Baichuan4-Turbo,
analysis: [
1. 科学家们通过分析阿波罗任务带回的月球岩石样本发现，月球上的火山活动确实在约30亿年前达到了高峰，之后逐渐减少。,
2. 迄今为止，尚未有确凿证据表明月球在30亿年后仍有活跃的火山活动。,
3. 通过对月球表面的观测和探测，包括撞击坑的密度和分布等特征，支持月球火山活动在数十亿年前就已经停止的观点。,
4. 月球的地质活动与地球相比非常微弱，其内部热量散失快，导致火山活动较早结束。,
5. 美国宇航局（NASA）和其他国际空间机构的研究均支持月球在远古时期存在火山活动，而这些活动在30亿年前已经基本结束。],
label_p: 0,
Error Type: E:Factual Inconsistency

English Example of Reasoning Inadequacy

id: 2128,
claim: Smelling sachets or essential oils can help improve reduced sense of smell after COVID-19 recovery.,
publish_date: 2023/1/18,
label: 0,
source: CCTV News Client,
gold evidence: Doctor suggested using strong-smelling items to stimulate the sense of smell.,
domain: Public Health,
model: GPT-3.5-Turbo,
analysis: [
1. Smelling sachets or essential oils lacks scientific evidence as a treatment for reduced sense of smell.,
2. Reduced sense of smell may indicate underlying pathological issues and requires professional diagnosis.,
3. Home remedies like sachets or essential oils might delay necessary medical treatment.,
4. Professional medical evaluation is critical for addressing symptoms effectively.,
5. Unverified remedies should be approached with caution to avoid potential health risks.],
label_p: 1,
Error Type: G:Under Informativeness

id: 28,
claim: Wuxi police recently arrested human traffickers.,
publish_date: 2024/10/15,
label: 1,
source: Chongqing Rumor Refutation,
gold evidence: Verified by police, no recent cases of child trafficking have occurred locally.,
domain: Society,
model: GPT-4o,
analysis: [
1. Wuxi police have a history of combating human trafficking and solving related cases.,
2. Combating human trafficking is a key priority for China's security agencies.,
3. China's Ministry of Public Security organizes nationwide operations against human trafficking.,
4. Media and police frequently report on human trafficking arrests, including in the Wuxi region.,
5. Human trafficking is a global problem, and China has implemented effective measures to address it.],
label_p: 0,
Error Type: F:Overgeneralized Reasoning

Chinese Example of Reasoning Inadequacy

id: 2128,
claim: 新冠康复后嗅觉减退可以通过闻香囊、精油等锻炼。,
publish_date: 2023/1/18,
label: 0,
source: 央视新闻客户端,
gold evidence: 张伯礼：嗅觉、味觉不好的时候，可以闻香囊、鲜花，还有精油、风油精等。可以闻一闻刺激性强的气味锻炼嗅觉。,
domain: 公卫,
model: GPT-3.5-Turbo,
analysis: [
1. 闻香囊、精油等并不能缓解嗅觉减退的症状。,
2. 嗅觉减退是一种病理状态，通常由多种原因引起，如感冒、鼻窦炎、脑部疾病等，需要及时就医进行诊断和治疗。,
3. 闻香囊、精油等虽然在一定程度上可以帮助人们放松情绪，但并没有科学证据表明它们可以治疗或改善嗅觉减退症状。,
4. 嗅觉减退通常需要通过专业的医学检查和治疗来解决，自行使用闻香囊、精油等方法不仅无效，还可能延误病情。,
5. 在出现嗅觉减退等健康问题时，应及时就医，接受专业医生的诊断和治疗，避免轻信未经科学验证的偏方或谣言。],
label_p: 1,
Error Type: G:Under Informativeness

id: 28,
claim: 巫溪警方最近抓捕了人贩子。,
publish_date: 2024/10/15,
label: 1,
source: 重庆辟谣,
gold evidence: 经巫溪警方核实，当地近期并未发生贩卖儿童案件，也未组织抓捕人贩子行动，网传"抓人贩子"系谣言。,
domain: 社会,
model: GPT-4o,
analysis: [
1. 巫溪警方在多次行动中积极打击人口贩卖犯罪，过去也曾多次成功破获相关案件。,
2. 中国公安机关一直将打击人口贩卖作为重要任务，并定期通报相关案件的抓捕情况。,
3. 根据中国公安部的公开信息，各地警方定期开展专项行动以打击人口拐卖犯罪。,
4. 媒体报道和警方通报中经常会有关于人口贩子的抓捕新闻，巫溪地区也不例外。,
5. 人口贩卖是全球性问题，中国政府和警方对此采取了积极措施，并取得了一定成效。],
label_p: 0,
Error Type: F:Overgeneralized Reasoning

1094