Verification with Transparency: The TrendFact Benchmark for Auditable Fact-Checking via Natural Language Explanation

Anonymous ACL submission

Abstract

While fact verification remains fundamental, explanation generation serves as a critical enabler for trustworthy fact-checking systems by producing interpretable rationales and facilitating comprehensive verification processes. However, current benchmarks exhibit critical limitations in three dimensions: (1) absence of explanatory annotations, (2) English-centric language bias, and (3) inadequate temporal relevance. To bridge these gaps, we present TrendFact, the first Chinese fact-checking benchmark incorporating structured natural language explanations. TrendFact comprises 7,643 carefully curated samples from trending social media content and professional fact-checking repositories, covering domains such as public health, political discourse, and economic claims. It supports various forms of reasoning, including numerical computation, logical reasoning, and common sense verification. The rigorous multistage construction process ensures high data quality and provides significant challenges. Furthermore, we propose the ECS to complement existing evaluation metrics. To establish effective baselines for TrendFact, we propose FactISR—a dual-component method integrating evidence triangulation and iterative self-reflection mechanism. Experimental results demonstrate that current leading reasoning models (e.g., DeepSeek-R1, o1) have significant limitations on TrendFact, underscoring the real-world challenges it presents. FactISR significantly enhances reasoning model performance, offering new insights for explainable and complex fact-checking.

1 Introduction

The proliferation of counterfeit claims poses significant societal risks, including mass panic, social destabilization, and even armed conflicts, as exemplified by the COVID-19 infodemic(van Der Linden et al., 2020; Aondover et al., 2024). This critical challenge has driven substantial research efforts in automated fact-checking systems, particularly in developing comprehensive benchmark datasets. The rapid expansion of open datasets has accelerated advancements in AI-powered verification technologies, especially through large language models(Atanasova, 2024; Rani et al., 2023; Wang and Shu, 2023; Bilal et al., 2024; Kao and Yen, 2024). Despite these developments, current fact-checking benchmarks exhibit several critical limitations:

First, as an emerging subtask in fact-checking, explanation generation plays a pivotal role in producing interpretable results and enabling a comprehensive fact-checking process. Existing fact-checking benchmarks primarily focus on fact verification and evidence retrieval, with minimal attention to textual explanations. Additionally, the current datasets with textual explanations are mostly generated by LLMs and serve as intermediate steps to enhance the verification process. They lack the human validation to ensure the explanations' quality. This limitation hinders the interpretability of fact-checking to some extent.

Second, existing fact-checking benchmarks such as Factcheck-Bench (Wang et al., 2024), Feverous (Aly et al., 2021), and QuanTemp (Venktesh et al., 2024) have primarily focused on English scenarios, and only a few studies focus on other languages, such as Chinese (Lin et al., 2024; Hu et al., 2022). In the vast influx of information on the internet, the Chinese language represents a crucial source, second only to English data. The lack of Chinese in fact-checking benchmarks significantly limits the comprehensive applicability of artificial intelligence in fully addressing real-world fact-checking scenarios. Moreover, these benchmarks pay limited attention to trending topics in real-time, which hinders the development of practical and trustworthy AI.

To that end, we present TrendFact, the first benchmark for fact-checking in Chinese scenarios that incorporates structured natural language explanations. It contains 7614 samples from multiple data sources and includes various reasoning types, including numerical computation, logical reasoning, and common-sense errors, covering extensive areas like health, politics, and economics. Additionally, we introduce a new metric, ECS (Explanation Consistency Score), to mitigate the shortcomings of previous textual metrics that were insufficient for fully evaluating explanations.

We conduct a rigorous and thorough benchmark construction process. Specifically, we first curate a large-scale set of real-world fact statements from multiple trending platforms and previous research, ensuring the practicality and diversity of the data. Subsequently, we conduct a rigorous filtering process to identify challenging samples. Moreover, we employ significant human efforts to rewrite the remaining samples to incorporate discernible factors. Subsequently, we make adjustments to the evidence and annotate the structured natural language explanations for these samples. Finally, we review the final dataset to mitigate potential biases.

To establish effective baselines for this benchmark, we propose FactISR (Augmenting **Fact**-Checking via **I**terative **S**elf-**R**eflection), a methodology that systematically combines reasoning adaptation through evidence triangulation with iterative self-reflection enabled by reward decoding. We evaluate TrendFact on five of the strongest existing reasoning LLMs and two current fact-checking methods. We observed that the leading reasoning models exhibit certain limitations in their performance on TrendFact, and our FactISR method shows outstanding improvements when applied to base reasoning models.

In summary, our contributions are as follows:

- We introduce TrendFact, a comprehensive and challenging Chinese fact-checking benchmark that includes natural language explanations. TrendFact integrates real-world trending events with domain-specific factual data, creating a benchmark for fact-checking. To the best of our knowledge, it is the first benchmark to address both fact verification and explanation generation tasks in non-English scenarios.

- We propose FactISR - a dual-component method integrating evidence triangulation and an iterative self-reflection mechanism.

- We propose the Explanation Consistency Score (ECS) to evaluate the factual accuracy and consistency between generated explanations and ground-truth explanations, thereby complementing existing evaluation methodologies.

- We conducted extensive experiments that reveal limitations in the capabilities of current leading models on TrendFact. Additionally, we compared the performance of existing fact-checking methods with FactISR. The results show that FactISR can significantly enhance the performance of reasoning models.

2 Related Work

Fact-checking Benchmarks Existing fact-checking benchmarks can be divided into two primary categories based on their data sources. The first category includes benchmarks derived from Wikipedia data, such as STATPROPS (Thorne and Vlachos, 2017), FEVEROUS (Aly et al., 2021), and Hover (Jiang et al., 2020). The second category focuses on datasets developed by refining knowledge bases from fact-checking websites and existing fact-checkers, such as CLAIMDECOMP(Chen et al., 2022), DECLARE(Popat et al., 2018), and QUANTEMP(Venktesh et al., 2024). Although these benchmarks are considered comprehensive and challenging, most of them focus primarily on claim design and evidence retrieval, overlooking an essential aspect of the fact-checking task: explanation generation. As large language models play an increasing role in fact-checking, particularly by generating explanations to aid in verification, there is an urgent need for a benchmark that incorporates explanations to assess the reliability of this process. Unfortunately, most existing benchmarks neglect this crucial component, resulting in incomplete evaluations. Furthermore, only a few benchmarks, such as X-Fact (Gupta and Srikumar, 2021), CFEVER (Lin et al., 2024), and CHEF (Hu et al., 2022), address the creation of fact-checking datasets for non-English scenarios. To effectively assess the reliability of rapidly evolving AI-driven fact-checking, the development of a non-English benchmark that includes explanations is crucial.

Automatic Fact-checking Research on automated fact-checking technologies can be broadly categorized into two areas: fact verification and

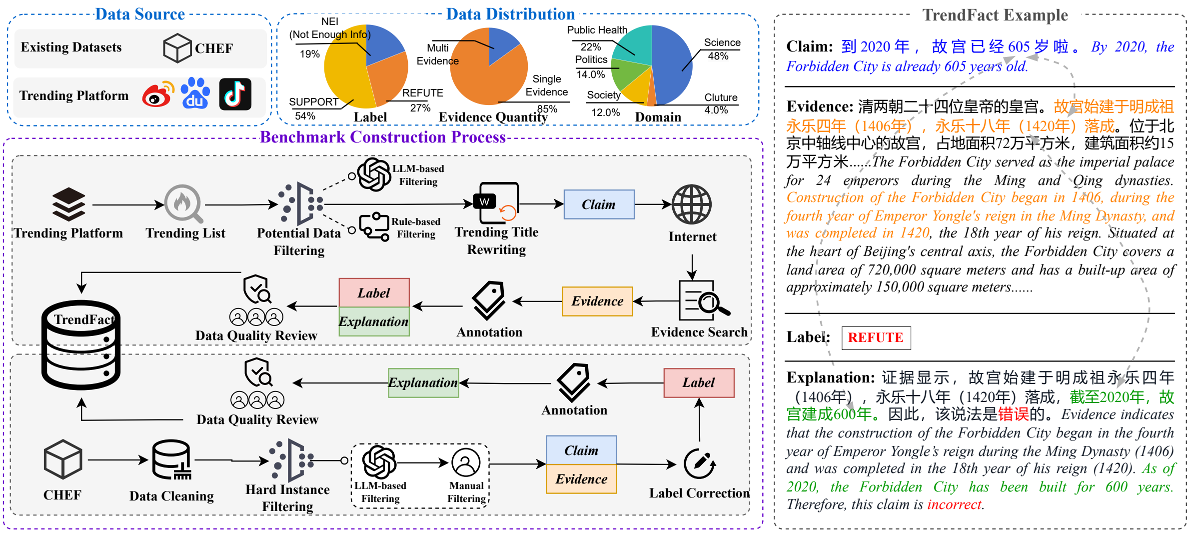

Figure 1: Overview of TrendFact. The left side illustrates the diverse data sources of TrendFact, the detailed distribution of the data, and the process of dataset construction. The right side displays a fact-checking example from TrendFact that involves complex numerical reasoning.

explanation generation. Fact verification is essential for the timely evaluation of claims. Previous studies have predominantly focused on areas such as Wikipedia article verification (Rashkin et al., 2017; Wang et al., 2022a), table-based verification (Liu et al., 2021; Zhou et al., 2022), and question-answering dialogue-based verification (Wang et al., 2022b; Zhang et al., 2024). With the advancement of large language models (LLMs), fact-checkers have begun leveraging these models to design fact-verification applications. For example, (Pan et al., 2023) proposes PROGRAMFC to utilize LLMs to generate executable programs, performing fact verification step by step. Explanation generation, on the other hand, addresses a more challenging task of producing interpretable results to support comprehensive fact-checking. It not only verifies claims but also provides an explanation of the principles behind the verification process. While most research has focused on using partial explanations as intermediate steps to facilitate verification, only a few studies have explored how natural language can be used to communicate the accuracy of claims and the rationale behind judgments. For example, (He et al., 2023) proposes to help users correct misinformation by generating counter-misinformation responses. Nevertheless, fact verification and explanation generation have largely overlooked the interdependent feedback loop between veracity and explanation, and most of the research has been limited to English-language scenarios.

3 Datasets Construction

3.1 Overview

We introduce TrendFact, a comprehensive and highly challenging fact-checking benchmark. Figure 1 provides an overview of the TrendFact pipeline, which includes data sources, benchmark construction process, data attribute distribution, and example data point. TrendFact consists of a total of 7,643 data entries collected from trending events and fact-checkers, encompassing categories such as digital computation, logical reasoning, and common-sense errors. The dataset is further divided into 1,131 multi-evidence entries and 6,512 single-evidence entries. Notably, TrendFact is a fact-checking benchmark to include verification explanations. In contrast to previous benchmarks, which primarily used Wikipedia and fact-checking websites as data sources, we have opted for more practical and real-world sources, including the trending platforms and the professionally rigorous CHEF dataset.

3.2 Dataset Construction

Data Collection We begin by identifying the sources of fact-checking data. Previous studies have primarily relied on Wikipedia or various fact-checking websites. For example, the CHEF dataset was originally compiled from fact-checking websites. To ensure alignment with existing research, we also considered gathering data from these same sources. As a result, we selected CHEF as one

Dataset	#Claims	Source	Explanation Focus	Language
Synthetic Claims				
FEVEROUS(Aly et al., 2021)	$87,026$	WP	$\times$	English
CHEF(Hu et al., 2022)	10000	FCS	$\times$	Chinese
Hover(Jiang et al., 2020)	$26,171$	WP	$\times$	English
CFEVER(Lin et al., 2024)	$30,012$	WP	$\times$	Chinese
STATPROPS(Thorne and Vlachos, 2017)	$4,225$	FB	$\times$	English
Fact-checker Claims				
CLAIMDECOMP(Chen et al., 2022)	$1,250$	Politifact	$\times$	English
DeClarE(Popat et al., 2018)	$13,525$	FCS	$\times$	English
X-Fact(Gupta and Srikumar, 2021)	$1,800$	FCS	$\times$	Multi
AVeriTeC(Schlichtkrull et al., 2024)	$4,568$	FCS	$\times$	English
FlawCheck(Kao and Yen, 2024)	$30,349$	FCS	✓	English
QUANTEMP(Venktesh et al., 2024)	$30,012$	FCS	$\times$	English
TrendFact	7643	TP	✓	Chinese

Table 1: Comparison of TableFact with other fact-checking datasets. In the table, WP refers to Wikipedia, FCS refers to fact-checking websites, TP refers to Trending Platform, and FB refers to FreeBase. By “Explanation Focus”, here we refer to dataset contains explanations.

of our key data sources, as it is one of the few datasets focused on the Chinese scenario. Furthermore, to diversify the sources in our benchmark, we collected data from multiple trending platforms from 2020 to 2024, thus broadening the dataset’s coverage to include a wider range of real-world fact-checking scenarios. In total, we gathered an initial set of approximately 500,000 claims.

Data Formalization Due to varying data sources, the initial data exhibits significant differences in completeness, with each dataset containing some level of interference noise. The CHEF dataset is relatively comprehensive, including claims, evidence, and judgment labels. However, these samples have a distinct characteristic: the evidence often directly incorporates the judgment label (e.g., "… is a rumor"), which reduces the challenge and introduces factual errors as noise. In contrast, the trending data consists solely of statements, which must be transformed into claims. Furthermore, not all trend data can be converted into challenging claims. To address these issues, we implemented a rigorous data enhancement process to reduce noise and improve data completeness.

Given the differences in properties between the CHEF and trend data, we applied separate enhancement processes to each source. For the CHEF data, the enhancement process involved data cleaning, selection of challenging samples, and label correction. Initially, we removed samples with significant noise in the claims. Then, using a combination of LLM voting and manual filtering, we identified more difficult samples. Afterward, we corrected factual errors and refined the original labels, resulting in a more polished version of the dataset, CHEF-EG, which consists of 1,131 claim-evidence-label triples. For the trend data, the enhancement process included data filtering, claim rewriting, and evidence retrieval. Initially, we evaluated the potential of the data points by combining LLM voting to assess their suitability as challenging fact-checking samples. This allowed us to filter out approximately 50,000 samples. Subsequently, we manually filtered the data and selected a few thousand samples. Since the original trend data could not be directly used as claims, we rewrote the selected samples to incorporate identifiable factors. Then, we manually retrieved evidence via Google to validate the truthfulness of the rewritten claims, resulting in a dataset consisting of 6,512 claim-evidence-label triples. Finally, we merge the 7643 claim-evidence-label triples from these two sources to form the intermediate state of TrendFact.

Explanation Generation Most existing fact-checking datasets lack explanation components, and to our knowledge, there is no fact-checking dataset with explanations for non-English scenarios, such as Chinese. With the advancement of

LLM technology, the role of explanation generation in fact-checking has grown significantly. It not only offers general users clear and intuitive reasoning but also allows fact-checkers to assess the accuracy of fact-checking methods based on explanation generation. Consequently, we annotated the intermediate TrendFact samples with detailed explanations, creating the final comprehensive fact-checking benchmark consisting of claim-evidence-label-explanation quadruples. Specifically, we recruited and formed an annotation team consisting of fourteen graduate students and one doctoral student. They all come from prestigious institutions and have undergone rigorous annotation training. Following the annotation process, we performed a quality review. Annotated data point was evaluated by three independent reviewers, in addition to the original annotator, to ensure high-quality annotations. This approach helps mitigate potential biases that may arise from individual annotators.

3.3 Comparisons with Existing Benchmarks

We perform a comparative analysis of TrendFact and existing fact-checking benchmarks, with the results summarized in Table 1. The results confirm our observation that current fact-checking benchmarks lack explanatory statements for fact verification. As a result, they fail to evaluate the explanation generation of fact-checking tasks, leading to an incomplete assessment. Additionally, most research has focused on English scenarios, with limited studies addressing non-English scenarios.

TrendFact, introduced in this work, is the first fact-checking benchmark for non-English scenarios that incorporates detailed explanations and integrates multiple data sources, including trending news websites. Unlike other benchmarks, TrendFact evaluates explanation generation in fact-checking tasks, providing greater practical relevance and broader coverage.

4 Baselines

In this section, we introduce the approaches used to comprehensively assess the difficulty and behaviors of TrendFact, including: LLMs, fact-checking methods, and our proposed method, FactISR.

4.1 Fact-checking Methods

A benchmark is typically designed to achieve two key objectives: effectiveness and challenging complexity. First, it should be compatible with existing fact-checking methods, demonstrating its validity and acting as a practical testbed for current technologies. Second, it must be sufficiently complex to expose the limitations of existing methods, thus emphasizing its challenging nature. Accordingly, based on these objectives, we selected two representative automated fact-checking methods, PROGRAMFC (Pan et al., 2023) and CLAIMDECOMP (Chen et al., 2022), to evaluate TrendFact. We performed generation tasks on the TrendFact dataset for PROGRAMFC's decomposition procedures and CLAIMDECOMP's yes/no sub-questions, using GPT-3.5, with other parameters aligned with the original study.

4.2 LLMs

Recent advancements in large language models (LLMs) have given rise to advanced inference-scaled models, such as DeepSeek-R1 2025. These models, also known as inference-optimized LLMs, leverage techniques like reinforcement learning and inference scaling laws to significantly enhance reasoning capabilities. Their superior inference abilities make them suitable for complex tasks, including fact-checking. In fact-checking tasks, these models can process and evaluate claims against evidence, providing veracity assessments and explanations. Then, these advanced inference-scaled LLMs can thus play a pivotal role in improving the accuracy and reliability of fact-checking processes. In this work, we choose the following LLMs for testing TrendFact: QwQ-32B-preview (qwe, 2024), o1-preview (OpenAI, 2024), and DeepSeek-R1. Additionally, we select the powerful general-purpose LLMs, including GPT-4 (OpenAI, 2024), DeepSeek-v3 (Guo et al., 2025), and Qwen-72B-Instruct (qwe, 2024), which do not have enhanced reasoning capabilities, to broaden the evaluation scope. These models have also shown impressive abilities in tackling complex tasks.

4.3 FactISR

Inspired by the remarkable performance of techniques like DeepSeek-R1 in reasoning to solve complex tasks, we propose FactISR, an iterative self-reflection reasoning enhancing method. Figure 2 gives an overview of FactISR. It presents an iterative reasoning prompt template based on three key fact-checking features and combined with a reward decoding mechanism to promote base reasoning LLM's performance on fact-checking tasks.

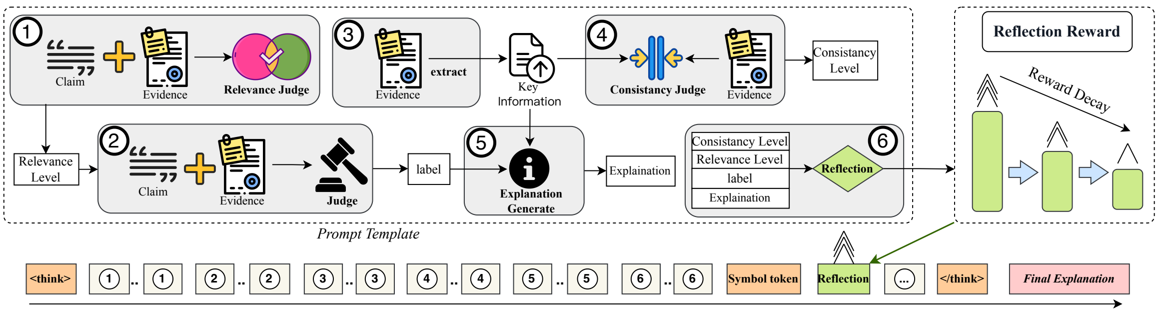

Figure 2: Overview of FactISR. The bottom section illustrates the model decoding process, where the <think> token signifies the commencement of the model's thought process according to the prompt template, concluding at </think>. On the right, it is indicated that the reward for reflection decreases incrementally with each instance of reflection. The Symbol token represents a sequence of tokens, upon hitting which a reflective token reward is provided for the generation of the subsequent token.

Reasoning Adaptation via Evidence Triangulation In order to enhance reasoning LLM, It is intuitive to construct a powerful prompt template. By analyzing the human iterative reasoning process on TrendFact, we identify three key features essential for fact-checking tasks: (1) the relevance between claim and evidence, (2) the consistency between key information and evidence, and (3) the conflict among claim, evidence, label, and explanation. First, we conclude that the higher the relevance, the greater the likelihood of ensuring the accuracy of the claim verification. Second, to ensure the correctness of extracted key information during the reasoning process of LLMs, they should perform self-feedback to validate the consistency between key information and evidence. Finally, LLMs must reflect on their reasoning process, and identify potential conflicts for iterative thinking.

Then, we perform the construction of the reasoning template combined with these features. As illustrated in Figure 2, it comprises six interconnected modules. Specifically, based on the relevance, an initial verification label is derived. Then, the LLM extracts key information from evidence, assesses its consistency, and uses it to reassess preliminary label while providing detailed explanations. Finally, the LLM reflects on the entire process, identifies conflicts, and provides feedback to guide further reasoning. Detailed prompt information is provided in the Appendix 3.

Iterative Self Reflection via Reward Decoding Ideally, the LLM would follow the template to conduct iterative reasoning. However, when conflicts arise clearly during reasoning, the model fails to acknowledge its errors, leading to premature termination of reasoning and preventing reflection. To address this, we propose a reward decoding mechanism to affect the LLM's generation process (Figure's 2 right side), improving the probability of it into an iterative reflection phase to resolve conflicts. Specifically, we predefined a target token sequence to serve as the trigger for the reward. Specifically, when the LLM generates tokens that meet this condition, the reward is activated, increasing the probability of the next token being "yes". This approach can be formalized as follows:

$$
\begin{aligned}
\boldsymbol{x}'_{h+k} &= \mathrm{RD}(\boldsymbol{x}_{h+k}) \\
&= \begin{cases} \boldsymbol{x}_{h+k} + \Delta_0 \cdot \gamma^i \cdot \boldsymbol{R}, & \text{if } \boldsymbol{x}_{h:h+k-1} = \boldsymbol{S} \\ \boldsymbol{x}_{h+k}, & \text{otherwise} \end{cases}
\end{aligned} \tag{1}
$$

Where $\boldsymbol{x}'_{h+k}$ represents the adjusted token at position $h+k$ in the LLM, while $\boldsymbol{x}_{h+k}$ is the originally generated token. $\boldsymbol{x}_{h:h+k-1}$ denotes the continuous token sequence generated by the LLM from position h to $h+k-1$, where $\boldsymbol{S}$ is a predefined specific token sequence. When $\boldsymbol{x}_{h:h+k-1}$ matches $\boldsymbol{S}$, we calculate the reward vector $\boldsymbol{R}$ (used to assign rewards to certain affirmative tokens) through the initial reward value Δ_0 and the decay factor γ (with the value range $0 < \gamma < 1$), with the i means iteration step.

5 Experiment

5.1 Experiment Setup

Metrics For fact verification, We choose four widely used and complementary metrics: F1-macro, Precision, Recall, and Accuracy. The F1-macro is the macro average of the three classes of F1 scores. Accuracy is the ratio of correctly verified samples to the total number of samples. For

Methods	Acc	F1	P	R
PROGRAM-FC	56.55	54.05	54.17	56.62
CLAIMDECOMP	59.35	56.86	56.65	59.41
Qwen-72B-instruct	65.14	60.56	66.97	63.65
DeepSeek-V3	63.74	60.31	66.09	63.96
GPT-4o	72.29	69.68	69.02	72.88
DeepSeek-R1	77.92	72.56	73.72	72.64
o1-preview	**78.98**	**75.16**	**75.13**	**75.72**
QwQ-32B-Preview	65.31	61.76	63.68	65.53
- w/ *COT*	71.90	68.64	68.21	71.25
- w/ *ET*	73.28	69.97	69.87	72.41
- w/ *ET+ISR*	75.45	72.44	72.48	74.59

Table 2: Comparison of FactISR with other baselines on fact verification task. The ET and ISR represent the evidence triangulation and iterative self-reflection components of the FactISR method, respectively. The last column presents the experimental results of the QwQ model after incorporating FactISR and COT.

the task of explanation generation, we employ the following evaluation metrics: BLEU-4, ROUGE (including ROUGE-1, ROUGE-2, and ROUGE-L), and BERTScore. Additionally, we introduce a new metric named explanation consistency score (ESC) to identify the consistency level between grounded explanation and generated explanation. The detail about ECS is presented in Appendix. These metrics were selected to comprehensively evaluate the textual similarity, word overlap, and fidelity of the generated explanations.

Baselines The methods for testing TrendFact have been detailed in **Baselines** section.

Experimental Settings. Our implementations are based on Pytorch. We leverage the QwQ-32B-Preview (hereafter referred to as QwQ) as the base LLMs. The evaluation for ESC was conducted using GPT-4o, while BERTScore evaluations are conducted on bert-base-chinese. The reward vector $\boldsymbol{R}$ and the decay factor γ are set to 20 and 0.1, respectively. The maximum input and maximum output length are set to 16k and 300, respectively. All inference experiments utilized greedy search as the strategy. All GPU-related inferences are executed on $4 \times A100$ GPUs.

5.2 Main Results

To provide a deeper insight into our proposed benchmark TrendFact and the fact-checking method FactISR, we present the evaluation results for fact verification and explanation generation in Table 2and Table 3, respectively. For fact verification, we conduct evaluation on three types of baselines. For the explanation generation, we select baselines exclude fact-checking methods, as these methods do not produce readable explanations.

Fact Verification Results For fact-checking methods, the results show that they perform poorly on the TrendFact benchmark. There's a big gap between existing fact-checking methods and others that are established on advanced-scaled inference LLM. This indicates that TrendFact contains many difficult samples and presents a high level of challenge to these methods. The diversity and complexity of the inference tasks within TrendFact are also the reasons for limiting their effectiveness. For LLMs, both general-purpose and advanced inference-enhanced LLMs, demonstrate superior performance, highlighting their robustness in addressing complex and challenging fact-chekcing problems. Moreover, advanced inference-enhanced LLMs outperform general-purpose models due to their superior reasoning capabilities. These capabilities allow them to tackle the complex, reasoning-based examples in TrendFact more effectively. Nevertheless, even the highest-performing model, o1-preview, failed to exceed 80% across all metrics. This also underscores the significant challenge posed by TrendFact.

Furthermore, compared to CoT (Chain-of-Thought)'s limited improvement, our proposed FactISR, which built on QwQ with two reasoning-improving strategies: Evidence Triangulation (ET) and Iterative Self Reflection (ISR), substantially enhanced the original QwQ's reasoning performance, with an improvement of 10.14 in accuracy and an F1 score increase of 10.68, achieving results on par with the second strongest models, DeepSeek-R1, while surpassing in recall metric. This confirms that our proposed method FactISRT is effective in enhancing LLMs with reasoning capabilities with handling more complex fact-checking tasks. In this study, we employ a rule-based parser to extract the response from the reasoning model's explanation, and then use [1] for obtaining the final veracity label.

Explanation Generation Results For LLMs, we observe that almost general-purpose LLMs universally outperform the advanced inference-scaled LLMs. For instance, o1-preview, which performs the best on the fact-verification task, lacks behind

[1]https://pypi.org/project/fuzzywuzzy/

Methods	BLEU-4	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	ECS
Qwen-72B-instruct	0.3366	0.8364	0.6441	0.4589	0.5906	0.7787
DeepSeek-V3	**0.3573**	**0.8432**	**0.6596**	**0.4805**	**0.6087**	0.7812
GPT-4o	0.2958	0.8270	0.6191	0.4189	0.5561	0.8622
DeepSeek-R1	0.2705	0.8143	0.5832	0.3821	0.5188	**0.9115**
o1-preview	0.2693	0.8022	0.5602	0.3960	0.5206	0.8986
QwQ-32B-Preview	0.2093	0.7804	0.5330	0.3459	0.4669	0.8198
- w/ *COT*	0.2790	0.8193	0.6012	0.4040	0.5443	0.8752
- w/ *ET*	0.2826	0.8057	0.5727	0.3927	0.5182	0.8861
- w/ *ET+ISR*	0.2815	0.8059	0.5727	0.3910	0.5173	0.8954

Table 3: Comparison of FactISR with other baselines on explanation generation.

the general-purpose LLMs on almost all metrics except for the ECS. Our analysis reveals that the reasoning process of inference-scaled models contains a significant amount of critical information, which leads to a clear reduction in the richness of their final outputs. These models tend to produce concise key pieces of information and accurate reasoning results rather than redundant non-critical content. In contrast, general-purpose LLMs excel at generating human-readable text, which produces more diverse and richer outputs. This results in higher scores on text-based metrics, such as 0.61 on ROUGE-L for DeepSeek-v3. However, since the ECS process includes key information from the reasoning model's inference process, these models can achieve outstanding performance on it. This phenomenon can be also validated through the ECS and the Acc in fact-verification: DeepSeek-R1 and o1-preview achieve higher ECS scores, indicating that their generated explanations are more consistent with the ground truth explanations and yield more precise results.

For FactISR, the results demonstrate its ability to significantly enhance the performance of the original reasoning model, QwQ, across all metrics. For example, the completed FactISR improves QwQ's score on BLEU-4 and ECS by 7.2 and 7.6, respectively. This improvement makes QwQ surpass the powerful reasoning model o1-preview in overall performance. This indicates that FactISR not only improves the reasoning accuracy of QwQ but also enriches its general-purpose output content. Compared to enhancements from CoT techniques, FactISR provides more substantial improvements in reasoning effectiveness.

5.3 Ablation Study

We evaluate the effectiveness of each component within FactISR by incrementally integrating them into QwQ. We conduct ablation experiments for both fact verification and explanation generation. The results (bottom of Table 2 and 3) demonstrate that adding the iterative reasoning template (ET) alone significantly improves the original QwQ's performance in both scenarios. For instance, QwQ's accuracy increased by 8 percentage points, surpassing GPT-4o, the best-performing general-purpose LLM in fact verification. Additionally, QwQ's BLEU-4 score exceeds that of DeepSeek-R1, the top-performing inference-enhanced LLM in explanation generation. When the iterative self-reflection module is incorporated, QwQ's performance further improves. These findings confirm the effectiveness of both FactISR components.

6 Conclusion

This paper introduces TrendFact, the first Chinese fact-checking benchmark with structured natural language explanations. Through a meticulously designed data construction process, TrendFact presents a diverse set of challenging samples requiring complex reasoning. We propose ECS to assess the factual accuracy and consistency between generated and ground-truth explanations. Experiments on automated fact-checking methods and advanced LLMs highlight TrendFact's considerable challenge. Additionally, we propose FactISR, a method that enhances base reasoning models, significantly improving their fact-checking performance.

7 Limitations

In this paper, we propose a Chinese fact-checking benchmark, TrendFact, which includes structured natural language explanations. However, to improve its real-time relevance, the claims in our dataset are sourced from trending statements on platforms, which require significant human effort to convert into more complex reasoning claims. Additionally, the evidence and explanations in the benchmark are manually gathered and summarized, resulting in high labor costs. We explore whether, in the future, more powerful LLMs with human-like summarization abilities can alleviate this issue.

References

2024. Qwen2 technical report.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Eric Msughter Aondover, Uchendu Chinelo Ebele, Timothy Ekeledirichukwu Onyejelem, and Omolara Oluwabusayo Akin-Odukoya. 2024. Propagation of false information on covid-19 among nigerians on social media. *LingLit Journal Scientific Journal for Linguistics and Literature*, 5(3):158–172.

Pepa Atanasova. 2024. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer.

Iman Munire Bilal, Preslav Nakov, Rob Procter, and Maria Liakata. 2024. Generating unsupervised abstractive explanations for rumour verification. *arXiv preprint arXiv:2401.12713*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*.

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*.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: a case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip S Yu. 2022. Chef: A pilot chinese dataset for evidence-based fact-checking. *arXiv preprint arXiv:2206.11863*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Wei-Yu Kao and An-Zi Yen. 2024. How we refute claims: Automatic fact-checking through flaw identification and explanation. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 758–761.

Ying-Jia Lin, Chun-Yi Lin, Chia-Jen Yeh, Yi-Ting Li, Yun-Yu Hu, Chih-Hao Hsu, Mei-Feng Lee, and Hung-Yu Kao. 2024. Cfever: A chinese fact extraction and verification dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18626–18634.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

OpenAI. 2024. Introducing openai o1-preview. Accessed: September 14, 2024.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744*.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

Anku Rani, SM Tonmoy, Dwip Dalal, Shreya Gautam, Megha Chakraborty, Aman Chadha, Amit Sheth, and Amitava Das. 2023. Factify-5wqa: 5w aspect-based fact verification through question answering. *arXiv preprint arXiv:2305.04329*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36.

James Thorne and Andreas Vlachos. 2017. An extensible framework for verification of numerical claims. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 37–40. Association for Computational Linguistics.

Sander van Der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about covid-19. *Frontiers in psychology*, page 2928.

V Venktesh, Abhijit Anand, Avishek Anand, and Vinay Setty. 2024. Quantemp: A real-world open-domain benchmark for fact-checking numerical claims. *arXiv preprint arxiv:2403.17169*.

H Wang and K Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. arxiv preprint arxiv: 231005253.

Hao Wang, Yangguang Li, Zhen Huang, and Yong Dou. 2022a. Imci: Integrate multi-view contextual information for fact extraction and verification. *arXiv preprint arXiv:2208.14001*.

Longzheng Wang, Peng Zhang, Xiaoyu Lu, Lei Zhang, Chaoyang Yan, and Chuang Zhang. 2022b. Qadialmoe: Question-answering dialogue based fact verification with mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3146–3159.

Yuxia Wang, Revanth Gangi Reddy, Zain Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230.

Xiaocheng Zhang, Chang Wang, Guoping Zhao, and Xiaohong Su. 2024. Li4: Label-infused iterative information interacting based fact verification in question-answering dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10488–10498.

Yuxuan Zhou, Xien Liu, Kaiyin Zhou, and Ji Wu. 2022. Table-based fact verification with self-adaptive mixture of experts. *arXiv preprint arXiv:2204.08753*.

A Details of the ECS

Table 4 provides a detailed definition of ECS. The descriptions from top to bottom are as follows: Dual Discrepancy: Both authenticity label misjudgment and fully inconsistent explanatory content. Label Error with Content Consistency: Incorrect authenticity labeling despite congruent explanatory material. Accurate Labeling with Explanatory Divergence: Correct authenticity identification accompanied by conflicting interpretation content. Partial Content Alignment: Proper authenticity classification with only partial consistency in explanatory elements. Full Verification Compliance: Complete congruence between correctly identified authenticity labels and their corresponding explanatory content.

B Prompt of ET

As shown in Figure 3, the prompt details the following aspects: (1)The relevance between the claim and evidence. (2)The consistency between key information and evidence. (3)The conflict among the claim, evidence, label, and explanation.

C Example of Rethinking via Reward Decoding

Figure 4 illustrates an example of rethinking via reward decoding. Without reward decoding, the model directly outputs a conclusion of no conflict and prematurely ends the reasoning process. Our reward decoding encourages the model to reassess its previous judgment, leading to a reconsideration that ultimately results in the correct outcome.

Label Verification Accuracy	Explanation Consistency	Score	Normalized Score
Misjudgment	Full Discrepancy	1	0.2
Misjudgment	Consistency	2	0.4
Correct Judgment	Content Divergence	3	0.6
Correct Judgment	Partial Consistency	4	0.8
Correct Judgment	Full Consistency	5	1.0

Table 4: Explanation consistency score.

你是一个事实核查专家，我将提供给你当前说法和相关证据，你要判断当前说法的真实性并给出一段文本解释。
当前说法的标签只有三种：真实，虚假，证据不充分。

你的输出应按如下形式：
如果说法真实：
<think>:{{思考过程在这里}}
<final_output>:证据显示，......。因此，该说法是正确的。
如果说法虚假：
<think>:{{思考过程在这里}}
<final_output>:证据显示，......。因此，该说法是错误的。
如果证据不充分：
<think>:{{思考过程在这里}}
<final_output>:证据显示，......。因此，证据不足以验证当前说法的真实性。

你可以选择的步骤有（相关性判断/真实性判断/关键信息提取/一致性判断/重新判断真实性/冲突性判断）
其中“冲突性判断”部分的输出话术为“我重新审视说法，证据，标签和解释之间是否可能存在冲突。*审视结果为：*（**否/是**）......”

你的思考过程应该参照如下形式（其中的‘/’为分隔符）：
<think>
第1步/相关性判断：我应该判断说法和证据之间的相关性，当前相关性判断是(高/中/低)。
第2步/真实性判断：我根据证据判断说法的真实性，由于相关度较*，并且***，(该说法应该是正确的。/该说法应该是错误的。/证据应该不足以验证当前说法的真实性。)
第3步/关键信息提取：我从证据中提取出关键信息，关键信息为“******”。
第4步/一致性判断：我判断关键信息与证据之间的一致性，当前一致性度量为（高/中/低）。
第5步/重新判断真实性：我根据关键信息反过头来验证之前判断的真实性，证据显示，......，因此，当前说法应该是**的，我之前的判断（准确/不够准确）。
第6步/冲突性判断（此步骤一定要严格按照给定话术进行输出）：我重新审视说法，证据，标签和解释之间是否可能存在冲突。审视结果为：**否**，思考结束/**是**(**和**可能存在冲突)，回到第*步重新思考。
（可选）第*步，******
</think>

注意：不能使用第一人称！
注意：你可以进行思考，但是最终的输出一定要严格按照给定格式！
再次强调：除<final_output>外的全部输出全部都要在<think>，即你的思考，最终的<final_output>应简洁规范！

You are a fact-checking expert. I will provide you with a current statement and related evidence, and you need to determine the truthfulness of the statement and provide an explanation.

The statement labels are limited to three categories: True, False, Insufficient Evidence.

Your output should follow this format:

If the statement is true:
<think>: { {Thought process here} }
<final_output>: The evidence indicates that... Therefore, the claim is true.

If the statement is false:
<think>: { {Thought process here} }
<final_output>: The evidence indicates that... Therefore, the claim is false.

If the evidence is insufficient:
<think>: { {Thought process here} }
<final_output>: The evidence indicates that... Therefore, the evidence is insufficient to verify the truthfulness of the claim.

Available steps include (Relevance Assessment/Truthfulness Assessment/Key Information Extraction/Consistency Assessment/Re-evaluation of Truthfulness/Conflict Assessment)
In "Conflict Assessment," the output should be: "Re-examining whether there is a potential conflict between the statement, evidence, label, and explanation. **The result is:** (**No/Yes**)..."

The thought process should follow this format (with '/' denoting separation):
<think>
Step 1/Relevance Assessment: Assess the relevance between the claim and the evidence. The current relevance is (high/medium/low).
Step 2/Truthfulness Assessment: Evaluate the truthfulness of the statement based on the evidence. Given the relevance is *, and ***, (the statement should be true./the statement should be false./the evidence should be insufficient to verify the statement's truthfulness.)
Step 3/Key Information Extraction: Extract key information from the evidence, which is "******".
Step 4/Consistency Assessment: Assess the consistency between key information and evidence. The current consistency measure is (high/medium/low).
Step 5/Re-evaluation of Truthfulness: Reassess the truthfulness based on key information. The evidence indicates that..., therefore, the statement should be **, and the previous assessment was (accurate/inaccurate).
Step 6/Conflict Assessment (this step must strictly follow the given phrasing): Re-examining whether there is a potential conflict between the statement, evidence, label, and explanation. **The result is:** **No**, thought process ends/**Yes** (** and ** may conflict), return to Step * for re-evaluation.
(Optional) Step *, ******
</think>

Note: First-person language should not be used!
Note: Thought processes should be enclosed within <think>, and the final <final_output> must be concise and standardized!
Emphasizing again: All outputs except <final_output> must be within <think>; the final <final_output> should be concise and formal!

Figure 3: The prompt template of this paper. The upper part is the Chinese prompt template, and the lower part is the translated English prompt template.

<think>
第1步/相关性判断：我需要判断说法和证据之间的相关性。说法提到2021年11月河南南阳市将迎来最强寒潮，最低气温降幅超10℃。证据中提到了11月6日至7日南阳市将有雨雪、大风、寒潮天气过程，最低气温降幅10℃～12℃，局部14℃～16℃。这与说法中的内容相符，因此相关性很高。
第2步/真实性判断：根据证据，南阳市确实将迎来寒潮，且最低气温降幅超过10℃，甚至局部地区超过14℃。证据来自中央气象台的预报，可信度较高。因此，初步判断该说法是真实的。
第3步/关键信息提取：从证据中提取关键信息，如寒潮的时间（11月6日至7日）、天气过程（雨雪、大风、寒潮）、气温降幅（10℃～12℃，局部14℃～16℃）等。
第4步/一致性判断：这些关键信息与说法中的“最强寒潮”和“最低气温降幅超10℃”一致，因此一致性很高。
第5步/重新判断真实性：基于上述一致性和证据的可靠性，我确认该说法是真实的。

No Reflection

第6步/冲突性判断：我重新审视说法、证据、标签和解释之间是否可能存在冲突。审视结果为：***否***，思考结束。
</think>

<final_output>
证据显示，2021年11月6日至7日，南阳市将经历雨雪、大风、寒潮天气，最低气温降幅达10℃～12℃，局部甚至超过14℃。这与说法中提到的“最强寒潮”和“最低气温降幅超10℃”相符。因此，该说法是正确的。
</final_output>

Reflection

第6步/冲突性判断：我重新审视说法、证据、标签和解释之间是否可能存在冲突。审视结果为：***是***，说法中提到“最强寒潮”，但证据中提到“全国自西北向东南，将有一次大风降温天气过程，大部地区气温将下降8℃～10℃，局地降16℃以上”，而南阳市的降幅为10℃～12℃，局部14℃～16℃。需要确认“最强寒潮”的表述是否准确。

第7步/进一步分析：为了验证“最强寒潮”的说法，需要了解此次寒潮是否为2021年最强。证据中提到这是“入秋以来最强寒潮”，但说法中没有限定“入秋以来”，只是说“最强寒潮”。此外，证据中提到大部地区降温8℃～10℃，局地16℃以上，而南阳市的降温幅度在10℃～12℃，局部14℃～16℃。因此，相对于大部地区，南阳市的降温幅度较大，但是否是全年最强，证据中没有提供足够的信息。

第8步/调整判断：由于证据中提到这是“入秋以来最强寒潮”，而说法中说的是“最强寒潮”，没有限定时间范围，因此可能存在不一致。此外，证据中没有提供全年寒潮的对比数据，无法确定这次是否是全年最强的寒潮。因此，说法可能过于绝对，证据不足以支持“最强寒潮”的说法。

第9步/最终判断：基于以上分析，证据支持南阳市将迎来一次强寒潮，气温下降幅度超过10℃，但是否是“最强寒潮”缺乏足够的信息支持。因此，说法的真实性无法完全确认，证据不充分。
</think>

<final_output>:
证据显示，南阳市在11月6日至7日将迎来一次寒潮天气，气温下降幅度超过10℃，局部地区甚至达到14℃～16℃。然而，说法中提到这是“最强寒潮”，而证据中仅描述为“入秋以来最强寒潮”，没有提供全年范围内的比较数据。因此，证据不足以验证当前说法的真实性。
</final_output>

Figure 4: An Example of Rethinking via Reward Decoding.