TrendFact: A Benchmark for Explainable Hotspot Perception in Fact-Checking with Natural Language Explanation

Anonymous ACL submission

Abstract

Although 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 have limitations that include the lack of impact assessment, insufficient high-quality explanatory annotations, and an English-centric bias. To address these, we introduce TrendFact, the first hotspot perception fact-checking benchmark that comprehensively evaluates fact verification, evidence retrieval, and explanation generation tasks. TrendFact consists of 7,643 carefully curated samples sourced from trending platforms and professional fact-checking datasets, as well as an evidence library of 66,217 entries with publication dates. We further propose two metrics, ECS and HCPI, to complement existing benchmarks by evaluating the system's explanation consistency and hotspot perception capability, respectively. Experimental results show that current fact-checking systems, including advanced RLMs such as DeepSeek-R1, face significant limitations when evaluated on TrendFact, highlighting the real-world challenges posed by it. To enhance the fact-checking capabilities of reasoning large language models (RLMs), we propose FactISR, which integrates dynamic evidence augmentation, evidence triangulation, and an iterative self-reflection mechanism. Accordingly, FactISR effectively improves RLM 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 (LLMs) (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 the field of fact-checking, explanation generation plays a pivotal role in enhancing the interpretability of verification results and refining the fact-checking process. However, existing benchmarks mostly lack textual explanations, making them inadequate to effectively evaluate the explanation generation performance. This limitation not only hinders the comprehensive evaluation of fact-checking but also restricts the improvement of the interpretability.

Second, with rapidly emerging information, the ability of fact-checking systems to perceive and check hotspot events has become increasingly important. However, existing benchmarks mostly focus on creating diverse and challenging samples while neglecting the hotspot characteristics of events. This renders them incapable of effectively measuring fact-checking systems' ability to respond to hotspot events. Moreover, the absence of Chinese benchmarks also hampers the comprehensive development of AI in the field of fact-checking.

To that end, we introduce TrendFact, the first hotspot perception fact-checking benchmark in Chinese scenario that incorporates a comprehensive evaluation of fact verification, evidence retrieval, and explanation generation tasks. It contains 7,643 samples collected from several trending platforms and existing datasets, as well as an evidence library of over 66,217 entries. Each sample in TrendFact consists of attributes including: claim, label, gold evidence, textual explanation, influence score, and four hotspot indicators. TrendFact dataset covers five domains such as healthcare and economy, as well as various reasoning types like numerical calculation and logical inference. Additionally, we propose two new metrics, ECS (Explanation Consistency Score) and HCPI (Hotspot Claim Perception Index), to support the evaluation of fact-checking systems on explanation generation and hotspot perception capability on TrendFact.

The construction of TrendFact consists of three components: (1) Data Collection; (2) Data Augmentation; (3) Evidence Library Construction. For the data collection, we collect a large-scale real-world fact entries from Weibo, Baidu and DouYin and existing dataset CHEF as initial data. For the data augmentation, we conduct two separate augmentation process for the initial data, including rigorous filtering process, claim rewrite, gold evidence retrieval, and textual explanation annotate. For the evidence library construction, we combine the gold evidence from dataset and the evidence retrieved from specifically defined process.

Additionally, We propose FactISR (Augmenting Fact-Checking via Iterative Self-Reflection), a methodology that enhances RLMs on fact-checking tasks by combining dynamic evidence augmentation, reasoning adaptation through evidence triangulation, and iterative self-reflection via reward decoding. We evaluate TrendFact on advanced LLMs, RLMs and fact-checking methods. The experiments results show that most test methods exhibit limited performance and our method FactISR introduce a improvement when applied to RLMs.

In summary, our contributions are as follows:

- We introduce TrendFact, a hotspot perception fact-checking benchmark in Chinese scenario that incorporates comprehensive evaluation on evidence retrieval, fact verification, and explanation generation tasks. To the best of our knowledge, it is the first benchmark that address the challenge of hostspot perception and explanation generation capability evaluation of fact-checking systems.

- We propose FactISR, which integrates dynamic evidence addition, evidence triangulation, and an iterative self-reflection mechanism to enhance the reasoning ability of RLMs on fact-checking task.

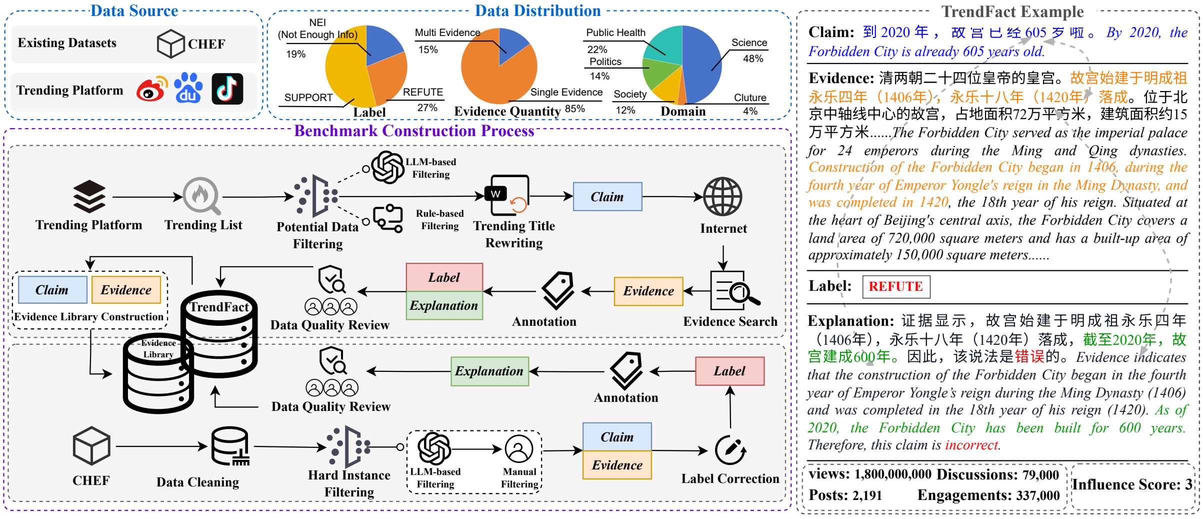
- We propose two new metrics ECS and HCPI to evaluate fact-checking systems on explanation generation and hotspot perception.

- We conduct extensive experiments that reveal limitations in the capabilities of LLMs, RLMs and fact-checking methods on TrendFact. Additionally, the results show that FactISR can enhance the performance of RLMs.

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). These benchmarks primarily focus on crafting datasets of diverse data types and challenging reasoning tasks and mainly target fact verification and evidence retrieval task, while overlooking the assessment of explanation generation. As LLMs play an increasing role in fact-checking, particularly by generating explanations to aid in verification, it is crucial to assess the reliability of this process. Additionally, evaluating the ability of fact-checking systems to perceive hotspot events is also a pressing issue in the field. However, most existing benchmarks neglect the two key needs above. Moreover, only a few benchmarks(Gupta and Srikumar, 2021; Lin et al., 2024; Hu et al., 2022) have paid attention to datasets in non-English contexts. Therefore, to comprehensively evaluate fact-checking systems, it is essential to develop a non-English benchmark that includes explanations and can assess the system's ability to perceive hotspots.

Automatic Fact-checking Research on automated fact-checking primarily falls into two categories: fact verification and explanation generation. Fact verification focuses on timely claim evaluation and has been widely explored in contexts such as Wikipedia articles, table-based data, and QA dialogues. With the rise of LLMs, methods like PROGRAMFC (Pan et al., 2023) generate



Claim: 到 2020 年，故宫已经 605 岁啦。*By 2020, the Forbidden City is already 605 years old.*

Evidence: 清两朝二十四位皇帝的皇宫。故宫始建于明成祖永乐四年（1406年），永乐十八年（1420年）落成。位于北京中轴线中心的故宫，占地面积72万平方米，建筑面积约15万平方米......*The Forbidden City served as the imperial palace for 24 emperors during the Ming and Qing dynasties. Construction of the Forbidden City began in 1406, during the fourth year of Emperor Yongle's reign in the Ming Dynasty, and was completed in 1420, the 18th year of his reign. Situated at the heart of Beijing's central axis, the Forbidden City covers a land area of 720,000 square meters and has a built-up area of approximately 150,000 square meters......*

Label: REFUTE

Explanation: 证据显示，故宫始建于明成祖永乐四年（1406年），永乐十八年（1420年）落成，截至2020年，故宫建成600年。因此，该说法是错误的。*Evidence indicates that the construction of the Forbidden City began in the fourth year of Emperor Yongle's reign during the Ming Dynasty (1406) and was completed in the 18th year of his reign (1420). As of 2020, the Forbidden City has been built for 600 years. Therefore, this claim is incorrect.*

views: 1,800,000,000 Discussions: 79,000
Posts: 2,191 Engagements: 337,000
Influence Score: 3

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 the benchmark construction. The right side displays a fact-checking example from TrendFact that involves numerical reasoning.

executable programs to support step-by-step verification. Explanation generation aims to produce interpretable outputs, but most work treats it as an intermediate step rather than a core objective. Few studies explore using natural language to convey both claim veracity and reasoning, which is critical for model interpretability and user understanding. For instance, He et al. (2023) generate counter-misinformation responses to correct false claims. Overall, current approaches often neglect the interplay between verification and explanation, limiting both transparency and user trust.

3 TrendFact Construction

In this section, we describe the fact-checking benchmark TrendFact which comprises 7,643 fact-checking samples along with an evidence library of 66,217 evidences. Figure 1 provides an overview of the dataset attributes, the benchmark construction process, and an sample from TrendFact. Next, we will introduce the data attributes and the construction process of TrendFact in detail.

3.1 Data Attributes

The TrendFact includes data in five domains, such as public health, science, society, politics and culture. These data are sourced from Trending Platforms (Weibo, Baidu, DouYin) and existing fact-checking dataset CHEF. Each sample in TrendFact consists of attributes including: claim, label, gold evidence, textual explanation, influence score, and four hotspot indicators. Depending on the number of gold evidence items, the dataset is divided into two categories: single-evidence and multi-evidence. Among these, samples with a single gold evidence account for 85%. For more details on hotspot indicators, importance scores, and other attributes, see Appendix A.

3.2 Benchmark Construction

We divide the construction process of TrendFact into three modules: data collection, data augmentation, and evidence library construction. Each module will be described in detail below.

3.2.1 Data Collection

In contrast to previous datasets, which primarily relied on data sources with limited timeliness such as Wikipedia and fact-checking websites, we select more practical and up-to-date sources that can capture emerging trends in a timely manner, including:

- **Trending Platforms.** Trending platforms such as Weibo, DouYin, and Baidu contain a wealth of dynamic and up-to-date factual content that can be leveraged for fact-checking. For example, their daily hotspot ranking pages often include concise factual statements accompanied by hotspot indicators such as view counts and discussion volumes. These indicators carry real-time dynamic features that are crucial for evaluating whether a fact-checking system possesses hotspot perception capabilities. Moreover, the content filtering mechanisms of trending platforms provide a certain

Dataset	#Claims	Source	Language	Explanation Focus	Hotspot Perception	Task		
						Evidence Retrieval	Claim Verification	Explanation Generation
Synthetic Claims								
FEVEROUS(Aly et al., 2021)	87,026	WP	English	×	×	Yes	Yes	No
CHEF(Hu et al., 2022)	10,000	FCS	Chinese	×	×	Yes	Yes	No
Hover(Jiang et al., 2020)	26,171	WP	English	×	×	Yes	Yes	No
CFEVER(Lin et al., 2024)	30,012	WP	Chinese	×	×	Yes	Yes	No
STATPROPS(Thorne and Vlachos, 2017)	4,225	FB	English	×	×	Yes	Yes	No
Fact-checker Claims								
CLAIMDECOMP(Chen et al., 2022)	1,250	Politifact	English	×	×	Yes	Yes	No
DeClarE(Popat et al., 2018)	13,525	FCS	English	×	×	Yes	Yes	No
X-Fact(Gupta and Srikumar, 2021)	1,800	FCS	Multi	×	×	Yes	Yes	No
AVeriTeC(Schlichtkrull et al., 2024)	4,568	FCS	English	×	×	Yes	Yes	No
FlawCheck(Kao and Yen, 2024)	30,349	FCS	English	✓	×	Yes	Yes	Yes
QUANTEMP(Venktesh et al., 2024)	30,012	FCS	English	×	×	Yes	Yes	No
TrendFact	7,643	TP	Chinese	✓	✓	Yes	Yes	Yes

Table 1: Comparison of TrendFact 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.

level of quality assurance for the factual information they present. Therefore, we collect approximately 500,000 hotspot event entries from these platforms between 2020 and 2024 as part of the initial dataset.

- **Existing Datasets.** Although the data sources of existing fact-checking datasets have dynamic limitations, they still offer valuable information for broadening verification scenarios beyond trending platforms. However, only a few datasets specifically focus on the Chinese scenario. For example, the CHEF dataset, which collects data from multiple fact-checking websites can serve as a representative resource. Therefore, we include CHEF as part of the data sources in our initial dataset.

3.2.2 Data Augmentation

The initial data collected from trending platforms and existing fact-checking datasets exhibit distinct limitations. First, the trending data lacks the key attributes required for serve as effective fact-checking samples and mostly consists of meaningless noise. Second, the evidence in the CHEF dataset often includes judgmental labels (e.g., "... is a rumor"), which reduces the difficulty of the fact-checking task. In addition, CHEF also contains noisy samples with factual inaccuracies. More information are detailed in Appendix B. Lastly, both data sources lack textual explanations, which are essential for TrendFact. To address these issues, we recruit an annotation team consisting of fourteen graduate students and one doctoral student to carry out a series of rigorous data augmentation processes for each of the two intial dataset.

Data from Trending Platforms We first employ a voting mechanism with multiple LLMs to filter out 90% of the meaningless samples and those lacking sufficient challenge potential. Subsequently, our annotation team further screens the remaining samples to remove sensitive or overlapping data, resulting in a final set of 6,512 samples. Since the initial trending data lacked the essential factual elements required for fact-checking claims, these samples could not be directly used as claims. To address this issue, we develop a rewriting guideline, and annotators manually rewrite the 6,512 samples accordingly. Consequently, for each of rewritten sample, we then manually retrieve one piece of gold evidence and provide a detailed explanatory annotation (more details in Appendix C).

Data from Existing Datasets For the CHEF dataset, we first manually filter out noisy data containing sensitive content and correct samples with factual errors. Then, similar to the trending data filtering, we apply an LLM voting mechanism to select samples that present significant verification challenges. Each of the remaining samples is then annotated with a detailed explanation by our annotation team. Consequently, the enhanced CHEF dataset comprises 1,131 high-quality samples.

After augmentation, we merge the two augmented datasets to construct the TrendFact dataset, which contains a total of 7,643 samples. Notably, we provide rigorous training to our annotation team

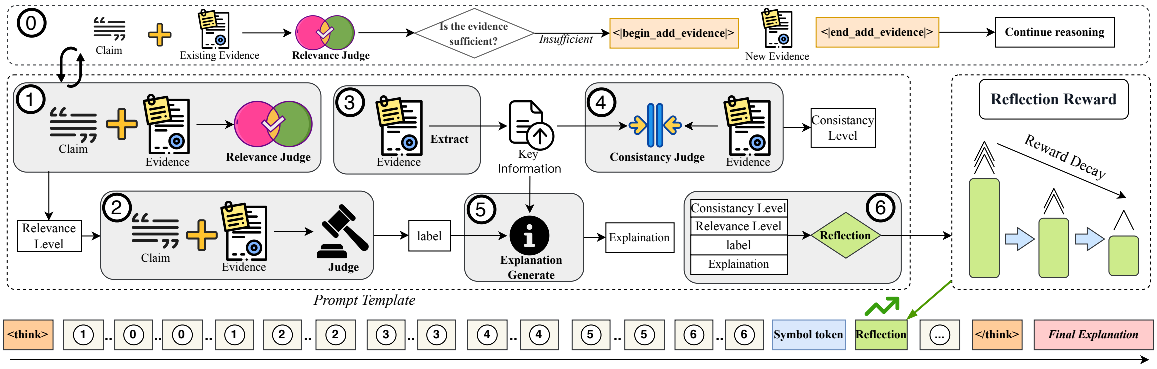


Figure 2: Overview of FactISR. The bottom section depicts the token generation process, where RLMs' thought is enclosed by <think> and </think>. When the "Symbol" token is hit, it provides a reward for the next token being a "Reflection" token. The topmost section shows DEA's dynamic evidence addition, while the middle-left details ET's reasoning steps. The middle-right illustrates the decreasing reward for "Reflection" tokens with increased reflections.

and conduct standardized human evaluations on both the rewritten and annotated results to ensure data quality. The evaluation criteria are detailed in Appendix D. While TrendFact primarily focuses evaluating the hotspot perception capabilities of fact-checking systems, we also aim to extend its utility to assess fundamental verification abilities. To this end, we set the hotspot indicator attribute to null for approximately 15% of the samples.

3.2.3 Evidence Library Construction

The Evidence Library is a crucial component of the fact-checking benchmark, providing essential support for accurate verification. Specifically, we first employ an LLM to generate a set of candidate claims that are semantically similar to the original claim but differ in content and structure. We then retrieve supporting evidence for these candidates from websites and combine it with the gold evidence from the TrendFact dataset to construct a comprehensive evidence library comprising 66,217 entries. The evidence library construction process is simply illustrated in Figure 1 and more detailed descriptions of the evidence library construction process is provided in Appendix E.

3.3 Comparisons with Existing Benchmarks

We conduct a comprehensive comparative analysis between TrendFact and existing fact-checking benchmarks, with the results summarized in Table 1. Our analysis reveals that most existing benchmarks fall short in evaluating explanation quality and hotspot perception capabilities, the two critical aspects for advancing the fact-checking field. These limitations significantly hinder the development of more robust and responsive verification systems. In contrast, TrendFact offers a comprehensive evaluation across three key fact-checking tasks and effectively assesses a system's ability to respond to emerging hotspot events.

4 FactISR

4.1 Overview

Inspired by the impressive reasoning capabilities of advanced RLMs like DeepSeek-R1 in complex tasks, we believe that they hold strong potential for addressing the challenges of fact-checking. However, there lack effective methods for conveying fact-checking task content to RLMs in a way that facilitates efficient reasoning. To bridge this gap, we propose FactISR, a framework designed to enhance the general fact-checking abilities of RLMs. An overview of FactISR is shown in Figure 2. It comprises three key components: dynamic evidence augmentation, reasoning adaptation via evidence triangulation, and iterative self-reflection via reward decoding. In the following sections, we introduce each of these components in detail.

4.2 Dynamic Evidence Augmentation

High-quality evidence is fundamental to accurate claim verification. It's an intuitive way to feed all retrieved evidence directly into the RLM to support the fact-checking. However, lengthy evidences often contain irrelevant or noisy information, which can hinder accurate judgment and lead to redundant reasoning time. To address these issues, we develop the Dynamic Evidence Augmentation (DEA) module. Specifically, DEA iteratively incorporates retrieved evidence into the RLM when specific signals are detected in the model's output. Figure 10 provides an example of the DEA.

4.3 Reasoning Adaptation via Evidence Triangulation

Given the augmented evidence, it remains challenging for RLMs to fully capture the relationships between the evidence and other key elements, such as the claim, explanation, and label, to make accurate judgments. To this end, we introduce a comprehensive reasoning template based on **Evidence Triangulation** to enhance the RLM's understanding and adaptation to fact-checking tasks.

Firstly, we construct the evidence triangulation framework based on the in-depth analysis of human reasoning patterns, derived from hundreds of verification actions performed by our annotation team. This framework comprises: (1) the relevance between the claim and evidence, (2) the consistency of key information extracted from evidence, and (3) the conflict within reasoning process. Then, we build a structured reasoning template based on these factors, as depicted in the six interconnected modules shown in Figure 2. Specifically, the RLM is first prompted to generate an initial verification label based on the relevance assessment. Next, it extracts key information from the evidence and evaluates its consistency with the claim, refining the initial label and producing an explanation. Finally, the RLM reflects on the entire reasoning process to identify any internal conflicts and generate feedback signals for further reasoning.

4.4 Iterative Self Reflection via Reward Decoding

Given the reasoning prompt, the ideal behavior for the RLM is to follow it and engage in iterative reasoning to eliminate conflicts within the reasoning process. However, we observe that the RLM occasionally fails to recognize conflicts, resulting in premature termination. To address this limitation and enhance reasoning continuity, we introduce a reward decoding mechanism that increases the likelihood of the RLM entering the iterative reflection phase, thereby resolving unresolved conflicts (as shown in the right part of Figure 2).

Specifically, we define the probability of the token "*yes*" as the reward signal and designate a specific token sequence near the end of the reasoning template as the reward trigger. Under this setting, the RLM automatically triggers the reward as reasoning nears completion, receiving a gradually decreasing probability for the next token being "*yes*". This guides RLM enters into iterative reasoning and allow it to reflect on previous conclusions to generate proper judgment. 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 RLM from position h to $h+k-1$. $\boldsymbol{S}$ is the 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 Setup

Metrics For evidence retrieval task, we choose R@k, where k=1,2,3,5. For verification task, we choose F1-macro, Precision, Recall, and Accuracy. For the explanation generation task, we first employ BLEU-4, ROUGE (including ROUGE-1, ROUGE-2, and ROUGE-L), and BERTScore to assess the quality of generated explanation. Then we introduce two novel metrics: ECS and HCPI. Specifically, ECS evaluates the alignment between the system's generated explanations and verification results, which jointly considers LLM-based explanation scoring and verification accuracy. HCPI integrates multiple attributes, including hotspot indicators, risk factor, and ECS, to assess a fact-checking system's ability to detect and respond to high-impact erroneous claims. HCPI encourages models to identify and tag high-impact errors, providing explanations for them. It also penalizes incorrect judgments of non-errors, enhancing applicability to real-world platforms. The detailed formulation information is provided in Appendix F.

Baselines In this work, we choose the following types methods as baseline, including RLMs, LLMs and existing fact-checking methods. For RLMs, we select the most advanced QwQ-32B, QwQ-32B-Preview (qwe, 2024), Qwen3-32B(Think) (Yang

Methods	R@1	R@2	R@3	R@5
BM25 -w/o *date*	35.92	48.11	56.54	65.94
BM25	36.70	49.07	57.28	66.75
text-emb-ada-002	28.35	37.42	42.56	61.93
bge-m3(dense)	**49.02**	**61.48**	**69.67**	**78.97**

Table 2: Experimental Results on Evidence Retrieval.

Methods	Acc	F1	P	R
PROGRAM-FC	45.24	44.28	43.34	45.30
CLAIMDECOMP	47.48	46.40	45.32	47.53
QwQ-32B-Preview	56.73	53.30	55.92	57.27
Qwen2.5-72B-instruct	58.64	51.96	58.90	56.28
Qwen3-32B(*No think*)	61.51	55.01	58.11	58.07
DeepSeek-V3	63.42	57.17	60.35	60.44
GPT-4o	62.29	59.45	60.54	63.05
Qwen3-32B(*Think*)	70.09	65.20	64.97	67.17
DeepSeek-R1	71.67	64.94	65.31	66.00
QwQ-32B	72.67	66.68	66.96	68.00
FactISR(*QwQ-32B*)	**74.32**	**67.58**	**69.08**	**68.06**
FactISR(*Qwen3-32B*)	72.46	65.01	66.07	65.29

Table 3: Comparison of FactISR with other baselines on fact verification task.

et al., 2025), and DeepSeek-R1 (Guo et al., 2025). For LLMs, we choose GPT-4o (OpenAI, 2024), DeepSeek-v3 (Liu et al., 2024), and Qwen2.5-72B-Instruct (qwe, 2024), Qwen3-32B(No Think). For fact-checking methods, we select two representative automated methods including PROGRAMFC (Pan et al., 2023) and CLAIMDECOMP (Chen et al., 2022). For retrieve methods, we choose the following advanced methods including BM-25(without date), BM-25, OpenAI's text-embedding-ada-002 and bge-m3(dense)(Chen et al., 2024).

Experimental Settings. We conduct experiments on pytorch and 2 × $A100$ GPUs. The evaluation for ESC was conducted using GPT-4o, while BERTScore evaluations are conducted on bert-base-chinese. More details are shown in Appendix J.

5.2 Main Results

Evidence Retrieval Results We evaluate the four retrieval methods on their ability to retrieve target gold evidence from the evidence library of TrendFact. As shown in Table 2, the strongest bge-m3 only achieves a moderate level of performance with R@5 less than 80%. This indicates that our unique evidence library construction is capable of collecting challenging evidence for distinguishing original gold evidence, thereby increasing the difficulty of TrendFact.

Fact Verification Results We perform a comprehensive evaluation of various methods on the verification task of TrendFact, including traditional automatic fact-checking methods, LLMs, and RLMs. The results are presented in Table 3, and can be summarized as follows: First, traditional fact-checking models perform the worst, with all accuracy scores falling below 50%. Second, both LLMs and RLMs achieve better performance, exceeding the 50% threshold. Notably, RLMs outperform LLMs, as they are better equipped to handle the complex reasoning required by many TrendFact samples. However, even the strongest RLM, QwQ-32B, fails to achieve an overall 70% verification performance on TrendFact underscoring the significant challenge posed by the high-quality and reasoning-intensive nature of TrendFact. Furthermore, FactISR contributes to an improvement in the verification performance of RLMs. For example, it improves the performance of the best performing models, QwQ-32B and Qwen3-32B, by 1.65% and 2.37% in accuracy. In addition, the enhanced QwQ-32B model achieves improvements across all metrics.

Explanation Generation Results We evaluate both LLMs and RLMs on the explanation generation task in TrendFact, focusing on three key aspects: explanation quality, ECS, and HCPI. The results are presented in Table 4. First, we observe that, despite superior performance in the verification task, RLMs generally generate lower-quality explanations compared to LLMs. For example, QwQ-32B, the best performer in verification, produces relatively low-quality explanations. We attribute this to the design of RLMs, which prioritizes concise key information and accurate conclusions over elaborative output, resulting in less detailed and rich explanations. Second, RLMs outperform LLMs in terms of explanation consistency, as measured by ECS, which is expected since ECS captures the alignment between explanations and the model's internal reasoning process. Furthermore, FactISR improves the explanation quality of RLMs without compromising consistency, enabling them to approach the explanation performance of LLMs. Finally, regarding hotspot perception capability as measured by HCPI, RLMs achieve the best results, and FactISR further enhances their perfor-

Methods	HCPI	ECS	BLEU-4	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L
QwQ-32B-Preview	0.4923	0.7689	0.1474	0.7479	0.4544	0.2702	0.3928
Qwen2.5-72B-instruct	0.5321	0.7193	**0.2994**	0.8163	**0.6000**	**0.4128**	**0.5446**
Qwen3-32B(*No think*)	0.5172	0.7587	0.2740	**0.8166**	0.5921	0.3949	0.5389
DeepSeek-V3	0.5718	0.7623	0.2609	0.8058	0.5711	0.3705	0.5182
GPT-4o	0.5655	0.7972	0.2351	0.7934	0.5456	0.3380	0.4794
Qwen3-32B(*Think*)	0.5679	0.8279	0.2378	0.7962	0.5475	0.3402	0.4897
DeepSeek-R1	0.6032	**0.8430**	0.2144	0.7833	0.5152	0.3111	0.4519
QwQ-32B	0.6110	0.8355	0.2214	0.7858	0.5237	0.3163	0.4622
FactISR(*QwQ-32B*)	**0.6336**	0.8375	0.2185	0.7866	0.5251	0.3198	0.4645
FactISR(*Qwen3-32B*)	0.6157	0.8268	0.2443	0.8015	0.5604	0.3585	0.5097

Table 4: Comparison of FactISR with Other Baselines on Explanation Generation.

Methods	Acc	ECS	HCPI
FactISR(*QwQ-32B*)	**74.32**	**0.8375**	**0.6336**
- w/o *DEA*	73.30	0.8342	0.6189
- w/o *ET*	73.16	0.8361	0.6081
- w/o *ISR*	73.46	0.8328	0.6146
FactISR(*Qwen3-32B*)	**72.46**	**0.8268**	**0.6157**
- w/o *DEA*	71.09	0.8168	0.5844
- w/o *ET*	70.24	0.8230	0.5605
- w/o *ISR*	71.25	0.8211	0.5935

Table 5: Ablation Study for Evaluating Each Component of method FactISR.

mance in this aspect. Strong hotspot perception requires checking system to consider various event attributes, such as hotspot indicators, importance, and judgment consistency, to effectively identify high-impact samples and reduce the risk of misinformation. Base LLMs often struggle to reason over such complex information, resulting in weaker perception capabilities. In contrast, RLMs leverage their reasoning strengths to capture implicit relationships among attributes to enable perception. Furthermore, FactISR further enhances it by guiding RLMs to reason more effectively.

5.3 Ablation Study

To evaluate the effectiveness of each component in FactISR, we conduct ablation studies by individually removing one of the three modules from the fully integrated QwQ-32B and Qwen3-32B. We assess performance using both the accuracy of the verification task and the ECS of the explanation generation task. The results are shown in Table 5. It demonstrates that removing any single module leads to performance degradation across both tasks, confirming the effectiveness of all components in FactISR. Furthermore, we conduct further ablation experiments to analyze the efficiency gains introduced by the DEA module. As shown in Table 6, DEA contributes to improved accuracy and substantial gains in reasoning efficiency, achieving average reductions of more than 50% in both reasoning time and length. These results indicate that DEA effectively mitigates the reasoning inefficiencies caused by lengthy evidence.

Methods	Acc	Time	Length
QwQ-32B	72.67	0.8321	2664
+ *DEA*	**73.16**	**0.4294** (↓48.40%)	**817** (↓69.33%)
Qwen3-32B	70.09	0.7850	2664
+ *DEA*	**70.24**	**0.3851** (↓50.94%)	**803** (↓69.86%)

Table 6: Impact of DEA on Per-Sample Generation Time and Input Evidence Length.

6 Conclusion

In this paper, we introduce TrendFact, the first hotspot perception fact-checking benchmark for evaluation of fact verification, evidence retrieval, and explanation generation tasks. It comprises 7,643 challenging samples and an evidence library containing 66,217 entities through a rigorous data construct process. We also propose two novel metrics, ECS and HCPI, to assess the explanation generation and hotspot perception capability of fact-chekcing systems. In addition, we present FactISR, a method designed to enhance the fact-checking capabilities of RLMs. Experimental results demonstrate that TrendFact poses challenges to numerous advanced LLMs, RLMs, and automated fact-checking methods, while FactISR effectively improves RLMs 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*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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*.

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.

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

A Details of Data Attributes

The hotspot attributes of TrendFact sample include: views, discussions, posts, and engagements. These indicators are crucial for assessing the hotspot perception capabilities of fact-checking systems. Specifically, views represent the number of times the sample has been viewed on sampled trending platforms; discussions indicate the number of times the sample has been discussed; posts refer to the number of posts triggered by the sample; engagements represent the number of users involved. Additionally, the influence score, which is assessed by an LLM to indicate the potential threat level if the claim were false, ranges from 1 to 5, with 5 being the highest threat. This score is also a key component in evaluating the hotspot perception capabilities of fact-checking systems.

B Data Cleaning and Hard Example Selection

Figure 3 demonstrates data cleaning examples from the CHEF dataset, primarily showing the removal of samples containing extensive garbled text in evidence sources. For filtering trending headlines with fact-checking potential from social platforms, this paper implements a progressive human-AI collaborative filtering strategy. The pipeline sequentially eliminates headlines at different stages: (1) Initial filtering using large language models (LLMs) with stage-specific prompts, followed by (2) manual verification when sample quantities become manageable. This multi-stage approach yields challenging yet verifiable candidate samples through layered refinement. Table 7 and Figure 4 respectively present examples of stage-specific prompts and the sample filtering workflow.

C Fact-Checking Claim Rewriting Factor

As shown in Table 8, six critical factors are proposed to systematically transform social media headlines into verifiable claims: (1) Temporal Anchoring converts vague temporal expressions into specific time references; (2) Data Granularity decomposes aggregated data into measurable units; (3) Ambiguity Resolution replaces probabilistic terms with deterministic statements; (4) Comparative Standard introduces quantifiable benchmarks; (5) Domain Knowledge integrates industry-specific parameters; and (6) Source Implication embeds provenance cues. These factors collectively enhance claim verifiability while preserving semantic coherence for automated processing. Concrete rewriting examples are illustrated in Figure .

D Annotate Evaluation Criteria

Our human evaluation criteria, as illustrated in Figure 6, are divided into three components: attribute-level review, fact verification review, and explanation generation review. Specifically, in the attribute-level review, we assess the individual quality and mutual consistency of each sample's claim, evidence, and explanation. For the fact verification review, annotators independently determine the veracity of each claim based on the corresponding evidence and compare their judgments with the labeled result. In the explanation generation review, we manually verify whether the annotated explanation meets the predefined standards, given the claim, evidence, and label of the sample.

E Evidence Library Construction Process

Figure 7 illustrates the detailed evidence library construction process for TrendFact. First, we employ GPT-4o to rewrite existing claims, with Figure 8 showcasing the specific rewriting prompt. Specifically, we transform original statements into challenging variants by maintaining core fact relevance (relevance score >0.6) and introducing at least three types of semantic alterations: subject displacement (e.g., institution → individual), causal inversion (e.g., "A leads to B" → "B triggers A"), degree conversion (e.g., "significant growth" → "slight fluctuation"), spatiotemporal shift (across years/regions), and quantification method change (absolute value → percentage), allowing for reasonable logical leaps. The rewritten claims are then processed through Bing web retrieval, retaining the top 10 search results and extracting their textual

Claim: 目前傳統市場的最大風險為人擠人，且未戴好口罩，因此，建議民眾分流上市場，一次多買，且減少去市場的次數。
Evidence: Amessagetoourcommunity.ReadourCOVID-19Responsehere.©Copyright2021InteracInteracandInterace-TransferareregisteredtrademarksandtheInteraclogo,InLifeandInLifedesignaretrademarksofInteracCorp.ThesecookiesareessentialforourSitestooperateproperly.EssentialcookiesensurebasicfunctionalitiesandsecurityfeaturesoftheSites.Thesecookiesdonotgatherpersonalinformationaboutyouthatcouldbeusedformarketingpurposes.ReportingandAnalyticsWeusereportingandanalyticstoolstocollectinformationabouthowvisitorsuseourSites.Thesetechnologiesprovideuswithanonymous,aggregateddata.AdvertisingWeusecookiesandothertrackingtechnologiestoprovideyouwithtailoredadswhileyoubrowsetheinternet.Todothis,wemightsharesomedatawiththirdpartyplatforms,includingbutnotlimitedtosocialmediaplatforms.

Claim: 2020年2月，武汉有疑似感染新冠病毒的病人无法就诊。
Evidence:このページではJavaScriptを使用しています。JavaScriptを有効にしてください。このホームページを、英語・中国語・韓国語へ機械的に自動翻訳します。以下の内容をご理解のうえ、ご利用いただきますようお願いします。PDFファイルを見るためには、AdobeReaderというソフトが必要です。AdobeReaderは無料で配布されていますので、こちらからダウンロードしてください。御意見募集やパブリックコメント国民参加の場法人番号6000012070001〒100-8916東京都千代田区霞が関1-2-2電話番号03-5253-1111（代表）

Claim: 2020年2月，武汉有疑似感染新冠病毒的病人无法就诊。
Evidence:注册|登录你好|注销网站支持IPV6首页\xa0>\xa0专题专栏\xa0>\xa0食品药品安全\xa0>\xa0消费警示相信大家都听说过花生有清除血管垃圾和软化血管的作用，这是真的吗？那是不是所有人都应该多吃花生？吃花生用不用去皮？下面就来说说\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0\xa0扫一扫在手机打开当前页扫一扫在手机打开当前页联系我们|使用帮助|加入收藏加入收藏|网站地图版权所有

Figure 3: Examples of CHEF Dataset Cleaning.

content. Subsequently, we merge and deduplicate the newly collected evidence with pre-existing gold evidence, followed by publication date crawling for each webpage to finalize the evidence library.

F Details of the ECS and HCPI

Table 9 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.

The formulaic definition of the influence score in the HCPI metric is as follows:

$$\mathrm{Inf}(i) = r_i \cdot \begin{pmatrix} \alpha \cdot \log(1+v_i) + \beta \cdot \log(1+d_i) \\ + \kappa \cdot \log(1+e_i) + \lambda \cdot \log(1+p_i) \end{pmatrix} \quad (2)$$

The influence score $\mathrm{Inf}(i)$ of the i-th sample is calculated first, where v_i, d_i, e_i and p_i represent the views, discussions, engagements, and posts of the i-th sample, respectively. α, β, κ and λ are the corresponding weights that sum up to 1. r_i represents the risk index determined by GPT-4o. After obtaining the influence score, we calculate the HCPI metric as follows:

$$\mathrm{HCPI} = \frac{\begin{array}{l} \displaystyle\sum_{i \in \mathrm{SUP}} \begin{cases} \mathrm{Inf}(i) \cdot \mathrm{ECS}(i), & \hat{y}_i = \mathrm{SUP} \\ -2 \cdot \mathrm{Inf}(i), & \hat{y}_i = \mathrm{REF} \end{cases} \\ \displaystyle + \sum_{j \in \mathrm{REF}} \mathrm{Inf}(j) \cdot \mathrm{ECS}(j) \cdot \delta(\hat{y}_j = \mathrm{REF}) \\ \displaystyle + \sum_{k \in \mathrm{NEI}} \begin{cases} -\mathrm{Inf}(k), & \hat{y}_k = \mathrm{REF} \\ \mathrm{Inf}(k) \cdot \mathrm{ECS}(k), & \hat{y}_k = \mathrm{NEI} \end{cases} \end{array}}{\displaystyle\sum_{m \in \text{All Claims}} \mathrm{Inf}(m)} \quad (3)$$

where SUP, REF, and NEI represent the three categories of fact-checking labels: Support, Refute, and Not Enough Information, respectively. ECS(i) denotes the Explanation Consistency Score for i-th sample, and m is the total number of samples.

G Prompt of FactISR

As shown in Figure 9, the prompt details the following aspects: 1.EDA: Detailed Process of Dynamic Evidence Addition 2.ET: (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.

H Example of Rethinking via Reward Decoding

Figure 11 illustrates an example of reflection via reward decoding. Without reward decoding, the

Original Trending Headlines	1	2	3	4	Manual Review
故宫今年600岁了 The Forbidden City is 600 years old this year					✓
跨年收视率 New Year's Eve viewership ratings	🚫	---	---	---	------
bilibili晚会 Bilibili Gala	🚫	---	---	---	------
90 后超六成压力来自房和车 60% of 90s-born face stress from housing and cars					✓
@#40####4%(^^)**6 @#40####4%(^^)**6	🚫	---	---	---	------
这是刻在骨子里的教养吧 This is deeply rooted upbringing, isn't it?		🚫	---	---	------
健康饮食秘诀揭秘 Secrets to Healthy Eating Revealed				🚫	------
工厂该怎么留住年轻人 How can factories retain young workers?		🚫	---	---	------
疫情啥时候能结束 When will the pandemic end?		🚫	---	---	------
敢于真实 做时间的朋友 Dare to be authentic, be a friend of time.			🚫	---	------
重庆加州花园 Chongqing California Gardens				🚫	------
央行降准0.5个百分点 Central bank cuts RRR by 0.5%.					🚫
2020 新规 2020 New Regulations				🚫	------
首批九零后30了 First 90s-born are 30 now.					🚫
2020有5个神奇的星期六 2020 has five magical Saturdays					🚫
四川自贡地震 Earthquake in Zigong, Sichuan				🚫	------
祝你新年快乐 Wishing you a Happy New Year			🚫	---	------
2023年首场流星雨 The first meteor shower of 2023			🚫	---	------

Figure 4: Examples of Progressively Staged Data Filtering Workflow for Fact-Checking Potential Data Selection.

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.

I Experiments Under Gold Evidence Conditions

Tables 10 and 11 present experimental results of fact verification and explanation generation tasks under gold evidence conditions for LLMs and RLMs. Since gold evidence was pre-defined (rendering the DEA module inapplicable), our FactISR method is excluded from this comparison. The results demonstrate significant improvements in fact verification metrics (accuracy: +5-10 percentage points; F1) and explanation generation quality. Specifically, the fact verification accuracy of these methods significantly improved by 5 to 10 percentage points, with DeepSeek-R1 and o1-preview achieving scores of 77.92% and 78.98%, respectively. Similarly, for the explanation generation task, DeepSeek-V3 achieved a BLEU-4 score of

Original Trending Headlines +	Date +	References →	Appropriate Label →	Rewriting as Claim
故宫今年600岁了 The Forbidden City is 600 years old this year.	2020-01-01	....	REFUTE	到2020年，故宫已经605岁啦 *By 2020, the Forbidden City is already 605 years old.*
直播业平均月薪9423元 The average monthly salary in the live streaming industry is ¥9,423.	2020-01-02	....	REFUTE	2019年三季度，直播的平均薪酬为9423元/月。除了主播，创意策划属性的视频策划、编剧、编导岗位薪酬也不赖，其中编导的招聘薪酬最高 In the third quarter of 2019, the average salary for live streaming was ¥9,423 per month. Besides streamers, positions such as video planning, scriptwriting, and directing, which require creative planning skills, also offered good salaries, with the recruiting salary for directors being the highest.
2019年全国楼市调控达620次 In 2019, the number of real estate market regulations nationwide reached as high as 620 times.	2020-01-03	....	REFUTE	2019年全国楼市调控次数近乎翻倍，人才政策发布较去年同期相比上涨超过40% In 2019, the frequency of real estate market regulations nationwide nearly doubled, and the issuance of talent policies increased by over 40% compared to the same period last year.
新冠病毒可存活5天由飞沫等传播 The novel coronavirus can survive for up to 5 days and is transmitted through droplets and other means.	2020-02-03	....	SUPPORT	新冠病毒可存活5天由飞沫，更多是通过手传播 The novel coronavirus can survive for up to 5 days and is transmitted via droplets, but is more commonly spread through contact with hands.
油价或现三连降 Oil prices may experience three consecutive decreases.	2024-05-26	....	NEI	截至2024年5月27日，国内油价调整共经历了“五涨三跌两搁浅”，92号汽油跌幅最大 As of May 27th, 2024, domestic oil price adjustments have experienced "five increases, three decreases, and two pauses," with 92-octane gasoline seeing the largest decline.

Figure 5: Examples of Rewriting Trending Headlines into Fact-Checkable Claims.

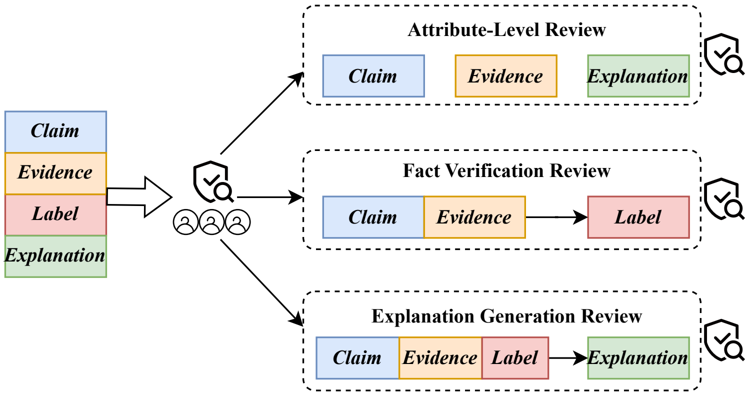


Figure 6: Annotate Evaluation Criteria.

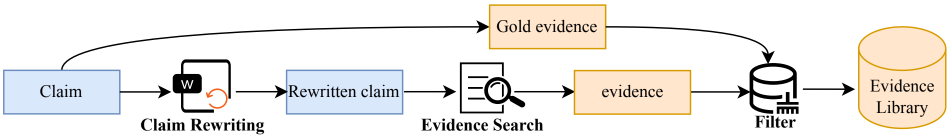


Figure 7: Evidence Library Construction Process.

0.3573, which is nearly 0.1 points higher than when using retrieval-based evidence.

J Details of the Experimental Settings

The reward vector $\boldsymbol{R}$ and the decay factor γ are set to 20 and 0.1, respectively. The maximum input length is set to 16k, while the maximum output lengths for LLMs and RLMs are set to 300 and 5k, respectively. The maximum length of retrieved results is truncated at 3k. The maximum number of retrievals is set to 3, and to ensure fair comparison, the maximum number of dynamically added evidences by our DEA is also limited to the same. All inference experiments utilized greedy search as the strategy. In this paper, for the HCPI metric, the values of α, β, κ and λ used to calculate the influence score are set to 0.05, 0.2, 0.15, and 0.6, respectively. Missing values are imputed using the 25th percentile, and the scores are scaled to ensure that the ratio of the maximum to the minimum influence score remains within a factor of 10. All GPU-related inferences are executed on $2 \times A100$ GPUs.

Iteration	Prompt
1	You are a trending topics analysis assistant, capable of accurately identifying the category of trending topics, mainly referring to trends on platforms like Weibo and Baidu. I will provide you with data on trending topics, and you need to help me determine their category. Note: I do not want entertainment-related trending topics. This means you do not need to output specific categories; you only need to decide whether a trending topic belongs to the entertainment category, and simply output one word: "Yes" or "No." Next, I will give you several examples for your reference in making judgments and outputs. **{Example Trending Topics}** Note: To emphasize again, you need to determine if a trending topic belongs to the entertainment category, and output only one word (Yes/No)! Note: If the trending topic is garbled text, also output No!
2	You are a trending topics analysis assistant, capable of accurately analyzing the category of trending topics, mainly referring to trends such as Weibo and Baidu hot searches. I will provide you with some trending topics data, and you need to help me determine whether these data are in question form. Note, you do not need to output specific categories; you only need to determine whether a trending topic is in question form and simply output one word: "Yes" or "No." Next, I will give you several examples for your reference to make judgments and outputs. **{Example Trending Topics}** Note: To emphasize again, you need to determine if the trending topic is in the form of a question and output only one word (Yes/No)! Note: Questions here may not necessarily contain a question mark or have obvious question features; they might be guiding sentences designed to attract clicks.
3	You are a fact-checking assistant, capable of accurately determining whether the current input can serve as a sample for fact-checking. It is known that a fact-checking task involves assessing the truthfulness of a claim based on provided evidence. However, I do not need to assess its truthfulness now; rather, I want to determine whether the current input has the potential to serve as a sample for a fact-checking dataset. I will provide you with real trending topics data from Weibo, and you need to help me assess whether these data have the potential to be included as samples in a fact-checking dataset. Note, you do not need to identify where the potential lies; you only need to output one word: "Yes" or "No." Next, I will give you several examples for your reference to make judgments and outputs. Examples are as follows: **{Example Trending Topics}** Note: You need to assess whether the trending topic has the potential to serve as a sample for a fact-checking dataset, and output only one word (Yes/No)! Note: Having potential means it contains elements that can be assessed and requires support from evidence, rather than abrupt statements or blessing words, etc.!
4	You are a fact-checking assistant, capable of accurately determining whether the current input can serve as a sample for fact-checking. It is known that a fact-checking task involves assessing the truthfulness of a claim based on provided evidence. However, I do not need to assess its truthfulness now; rather, I want to determine whether the current input has the potential to serve as a sample for fact-checking. More specifically, if the current input is merely in noun form, then it does not have the potential to be included as a sample in a fact-checking dataset. I will provide you with real trending topics data from Weibo, and you need to help me assess whether these data have the potential to be included as samples in a fact-checking dataset. Note, you do not need to identify where the potential lies; you only need to output one word: "Yes" or "No." Next, I will give you several examples for your reference to make judgments and outputs. Examples are as follows: **{More Challenging Trending Topic Examples}** Note: To emphasize again, you need to assess whether the trending topic has the potential to serve as a sample for a fact-checking dataset, and output only one word (Yes/No)! Note: Having potential means it is not merely a noun and contains elements that can be assessed, requiring support from evidence, rather than abrupt statements or blessing words, etc.!

Table 7: Progressively Staged Prompts for Fact-Checking Potential Selection.

Factor Category	**Definition**	**Rewriting Mechanism**
Temporal Anchoring	Adding/specifying temporal reference	Transforming vague temporal expressions into specific time nodes
Data Granularity	Disaggregating composite data into verifiable units	Decomposing aggregated data into independently verifiable dimensions
Ambiguity Resolution	Eliminating probabilistic/uncertain expressions	Replacing fuzzy quantifiers with deterministic statements
Comparative Standard	Establishing quantifiable reference standards	Introducing quantified comparison objects and proportions
Domain Knowledge	Incorporating professional contextual information	Supplementing industry-specific parameters or mechanisms
Source Implication	Indirectly indicating information provenance	Using industry-characteristic expressions to imply data sources

Table 8: Fact-Checking Claim Rewriting Factor

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 9: Explanation Consistency Score.

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
QwQ-32B-Preview	65.31	61.76	63.68	65.53
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**

Table 10: Experimental Results of Baselines Under Gold Evidence Conditions in Fact Verification Task.

Claim Rewriting Prompt for Evidence Retrieval

In-Depth Claim Rewriting Guidelines for Fact-Checking Systems

Core Objective

Generate challenging adversarial variants through transformative claim rewriting, requiring:

1. Maintain core factual relevance (semantic similarity score > 0.6)
2. Incorporate at least three semantic transformations from:
    - Subject substitution (e.g., organization → individual)
    - Causal inversion (e.g., "A causes B" → "B triggers A")
    - Magnitude alteration (e.g., "significant increase" → "minor fluctuation")
    - Spatiotemporal shift (across years/regions)
    - Quantification method change (absolute numbers → percentages)
3. Permit reasonable logical leaps

Transformation Strategies

- *Subject Generalization*: "Tesla brake incidents" → "Safety defects in an EV manufacturer" (Difficulty: Medium)
- *Temporal Ambiguity*: "Q2 2023" → "Recent summer seasons" (Difficulty: Low)
- *Data Recontextualization*: "30% growth" → "Falling short of projections" (Difficulty: High)
- *Causal Restructuring*: "Smoking causes lung cancer" → "Lung cancer patients frequently have smoking history" (Difficulty: Very High)
- *Composite Transformation*: "20% PM2.5 reduction in Beijing 2023" → "Northern China's air quality improvements failed to meet pledged targets" (Difficulty: Expert)

Input-Output Demonstration

Input:

"Tesla Model 3 sales in China increased 45% YoY in 2023, accounting for 18% of total NEV sales"

Outputs:

1. "A foreign EV brand's China market share exceeded 15% last year despite growth rates below industry expectations"
2. "North China sales records suggest potential discrepancies in delivery figures for a popular EV model during 2022-2023"
3. "Industry sources indicate leading EV manufacturers achieved annual sales growth primarily through price reductions"

Implementation Requirements

Generate ONE adversarial paraphrase that:

- Modifies ≥3 critical elements
- Employs distinct transformation strategies
- Maintains surface plausibility
- Avoids explicit factual errors
- Output format: Modified claim only (omit transformation labels)

Figure 8: Claim Rewriting Prompt for Evidence Retrieval.

Prompt of FactISR

Task Description

As a fact-checking expert, you are required to evaluate the veracity of a given claim based on provided evidence and generate a textual explanation. The claim can only be labeled as **True**, **False**, or **Insufficient Evidence**.

Methodological Constraints

1. **Dynamic Evidence Addition**

   Claims must be evaluated strictly based on evidence. If initial evidence sufficiently supports a **True** or **False** judgment, no additional evidence is required. If uncertain, generate the token `<|begin_add_evidence|>` to request external evidence.

   A maximum of **two** external evidence additions are permitted. Exceeding this limit triggers a system alert ("Maximum evidence threshold reached"), requiring judgment based on existing information.

   Evidence requests (e.g., "Current evidence is insufficient; additional evidence is required for validation. `<|begin_add_evidence|>`") must occur **before** the final explanation (i.e., within the reasoning phase).

2. **Evidence Triangulation Adaptation**

   Follow this structured reasoning workflow (*internal process only*):

   Step 1/Relevance Assessment: Evaluate claim-evidence relevance (High/Medium/Low).

   Step 2/Veracity Judgment: Determine claim veracity (e.g., "Given *[relevance level]* and *[evidence summary]*, the claim is likely **True/False**").

   Step 3/Key Information Extraction: Extract critical evidence snippets (e.g., "Key information: '***'").

   Step 4/Consistency Judgment: Assess consistency between key information and evidence (High/Medium/Low).

   Step 5/Re-evaluation: Reassess veracity using refined evidence (e.g., "Evidence demonstrates [...] Therefore, the claim is **False**").

   Step 6/Conflict Analysis (*strictly follow phrasing*):

   "Re-examining potential conflicts among the claim, evidence, and label. Result: **No conflicts**; final label confirmed as *[label]*."

   "Re-examining potential conflicts [...] Result: **Conflict detected** (*[specific conflict]*). Revisit Step *[X]*."

Guidelines

Avoid first-person pronouns.

Adhere strictly to the output format.

Final explanations must be concise (≤150 words, max 300 words).

Execute reasoning steps (including evidence addition and triangulation) **before** generating the final explanation.

Output Format

1. **True**:

   "The evidence indicates [...] Therefore, the claim is **substantiated**."

2. **False**:

   "The evidence indicates [...] Therefore, the claim is **refuted**."

3. **Insufficient Evidence**:

   "The evidence indicates [...] Therefore, the claim **cannot be verified** due to insufficient evidence."

Figure 9: English-Translated FactISR Prompt (partial formatting directives omitted for space constraints).

Claim: In China, crimes committed more than 20 years ago cannot be prosecuted under criminal proceedings.

<think>

Alright, I need to assess the veracity of the statement "In China, crimes committed more than 20 years ago cannot be prosecuted under criminal proceedings." First, examining the initial evidence, the user-provided Supreme People's Procuratorate report primarily discusses trends in criminal activity from 1999 to 2019, such as decreases in violent crime and increases in economic crime, but does not directly address the statute of limitations for prosecution. Therefore, the existing evidence does not cover whether crimes committed over 20 years ago are exempt from prosecution, necessitating further investigation into relevant legal provisions. ***Due to insufficient current evidence, additional evidence is required for verification.***

<|begin_add_evidence|>......<|end_add_evidence|>

Alright, I now have an additional piece of evidence, which I'll evaluate alongside the previous content to determine if it can verify the current claim. The new evidence is from the 2020 Amendment (XI) to the Criminal Law, which primarily deals with revisions concerning the age of criminal responsibility and hindrance of infectious disease prevention, but does not directly specify the statute of limitations period for prosecution. Therefore, further legal text investigation is needed to address the statute of limitations issue. ***Current evidence remains insufficient, necessitating additional evidence for verification.***

<|begin_add_evidence|>......<|end_add_evidence|>

Alright, I now have another piece of evidence, and I'll evaluate it along with the previous content to determine if it can verify the claim's accuracy. The new data from the Ministry of Public Security shows a decline in crime rates but does not address specific provisions concerning the statute of limitations. ***With evidence added twice, further additions are not possible.*** According to Article 87 of the Chinese Criminal Law, the statute of limitations varies by punishment type, with a maximum of 20 years, but if the crime is not prosecuted within this period without special circumstances, it may no longer be pursued. The claim that "after 20 years, crimes cannot be prosecuted" is overly absolute. Article 88 states that if the crime is discovered after the limitation period has passed, the prosecution may be initiated by the procuratorate. Therefore, the existing evidence indicates that the statement is not entirely accurate.

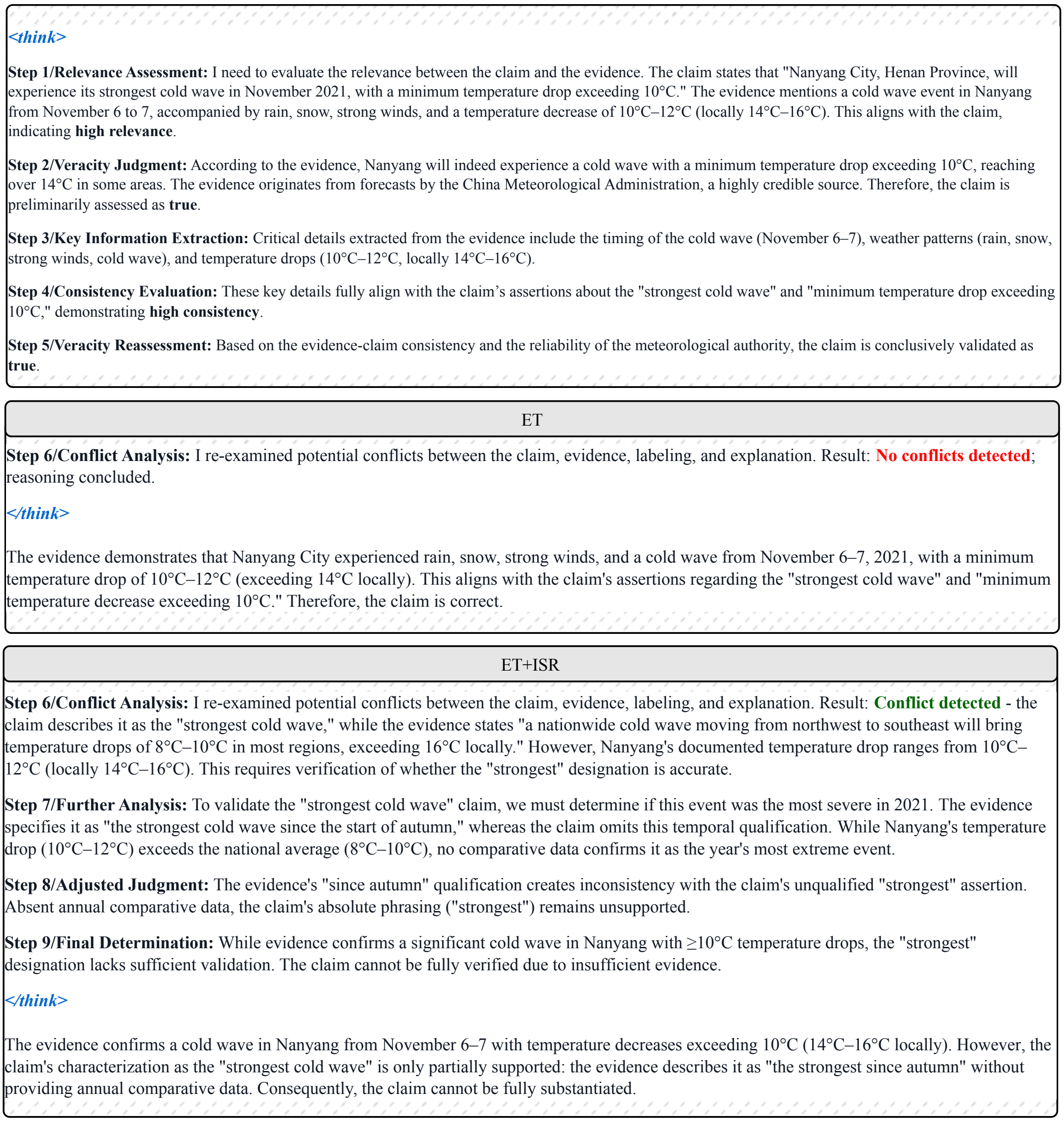
</think>

The evidence indicates that according to Chinese Criminal Law, the statute of limitations for prosecuting crimes varies depending on the type of punishment, with a maximum of 20 years. However, if a crime is discovered after the limitation period has expired, the procuratorate can still prosecute according to the law. Therefore, the claim is incorrect.

Figure 10: Example of Dynamic Evidence Addition.

Methods	BLEU-4	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L	ECS
QwQ-32B-Preview	0.2093	0.7804	0.5330	0.3459	0.4669	0.8198
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

Table 11: Experimental Results of Baselines Under Gold Evidence Conditions in Explanation Generation Task.



<think>

Step 1/Relevance Assessment: I need to evaluate the relevance between the claim and the evidence. The claim states that "Nanyang City, Henan Province, will experience its strongest cold wave in November 2021, with a minimum temperature drop exceeding 10°C." The evidence mentions a cold wave event in Nanyang from November 6 to 7, accompanied by rain, snow, strong winds, and a temperature decrease of 10°C–12°C (locally 14°C–16°C). This aligns with the claim, indicating **high relevance**.

Step 2/Veracity Judgment: According to the evidence, Nanyang will indeed experience a cold wave with a minimum temperature drop exceeding 10°C, reaching over 14°C in some areas. The evidence originates from forecasts by the China Meteorological Administration, a highly credible source. Therefore, the claim is preliminarily assessed as **true**.

Step 3/Key Information Extraction: Critical details extracted from the evidence include the timing of the cold wave (November 6–7), weather patterns (rain, snow, strong winds, cold wave), and temperature drops (10°C–12°C, locally 14°C–16°C).

Step 4/Consistency Evaluation: These key details fully align with the claim's assertions about the "strongest cold wave" and "minimum temperature drop exceeding 10°C," demonstrating **high consistency**.

Step 5/Veracity Reassessment: Based on the evidence-claim consistency and the reliability of the meteorological authority, the claim is conclusively validated as **true**.

ET

Step 6/Conflict Analysis: I re-examined potential conflicts between the claim, evidence, labeling, and explanation. Result: **No conflicts detected**; reasoning concluded.

</think>

The evidence demonstrates that Nanyang City experienced rain, snow, strong winds, and a cold wave from November 6–7, 2021, with a minimum temperature drop of 10°C–12°C (exceeding 14°C locally). This aligns with the claim's assertions regarding the "strongest cold wave" and "minimum temperature decrease exceeding 10°C." Therefore, the claim is correct.

ET+ISR

Step 6/Conflict Analysis: I re-examined potential conflicts between the claim, evidence, labeling, and explanation. Result: **Conflict detected** - the claim describes it as the "strongest cold wave," while the evidence states "a nationwide cold wave moving from northwest to southeast will bring temperature drops of 8°C–10°C in most regions, exceeding 16°C locally." However, Nanyang's documented temperature drop ranges from 10°C–12°C (locally 14°C–16°C). This requires verification of whether the "strongest" designation is accurate.

Step 7/Further Analysis: To validate the "strongest cold wave" claim, we must determine if this event was the most severe in 2021. The evidence specifies it as "the strongest cold wave since the start of autumn," whereas the claim omits this temporal qualification. While Nanyang's temperature drop (10°C–12°C) exceeds the national average (8°C–10°C), no comparative data confirms it as the year's most extreme event.

Step 8/Adjusted Judgment: The evidence's "since autumn" qualification creates inconsistency with the claim's unqualified "strongest" assertion. Absent annual comparative data, the claim's absolute phrasing ("strongest") remains unsupported.

Step 9/Final Determination: While evidence confirms a significant cold wave in Nanyang with ≥10°C temperature drops, the "strongest" designation lacks sufficient validation. The claim cannot be fully verified due to insufficient evidence.

</think>

The evidence confirms a cold wave in Nanyang from November 6–7 with temperature decreases exceeding 10°C (14°C–16°C locally). However, the claim's characterization as the "strongest cold wave" is only partially supported: the evidence describes it as "the strongest since autumn" without providing annual comparative data. Consequently, the claim cannot be fully substantiated.

Figure 11: An Example of Reflection via Reward Decoding.