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

Anonymous ACL submission

Abstract

Fact-checking benchmarks provide standardized testing criteria for automated fact-checking systems, driving technological advancement. With the surge of misinformation on social media and the emergence of various fact-checking methods, public concern about the transparency of automated systems and the accuracy of fact-checking for high infulence events has grown. However, existing benchmarks fail to meet these urgent needs and are predominantly English-centric, hindering the progress of comprehensive fact-checking. To address these issues, we introduce TrendFact, the first benchmark capable of evaluating hotspot perception ability (HPA) and all fact-checking tasks. TrendFact consists of 7,643 curated samples sourced from trending platforms and professional fact-checking datasets, as well as an evidence library containing 366,634 entries with publication dates. Additionally, to complement existing benchmarks in evaluating system explanation consistency and HPA, we propose two new metrics: ECS and HCPI. Experimental results show that current fact-checking systems face significant limitations when evaluated on TrendFact, which facilitates the development of more robust fact-checking methods. Furthermore, to enhance the capabilities of existing advanced fact-checking systems, the reasoning large language models (RLMs), we propose FactISR, a reasoning framework that integrates dynamic evidence augmentation with influence score-based iterative self-reflection. FactISR effectively improves RLM's performance, offering new insights into explainable and complex fact-checking.

1 Introduction

The proliferation of counterfeit claims poses significant social 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 drives substantial research efforts in automated fact-checking systems, particularly in developing comprehensive fact-checking benchmarks. The rapid expansion of open datasets has accelerated advancements in AI-powered verification technologies, especially through large language models (LLMs) and reasoning large language models (RLMs) (Atanasova, 2024; Rani et al., 2023; Wang and Shu, 2023; Bilal et al., 2024; Kao and Yen, 2024). Despite these developments, current benchmarks exhibit two key 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 ef-

Figure 1: A fact-checking example from TrendFact that involves numerical reasoning.

fectively evaluate the explanation generation performance. As our overview in Table 1 illustrates, this limitation not only hinders the comprehensive evaluation of fact-checking but also restricts the improvement of interpretability.

Second, different claims have varying degrees of social influence and should receive different levels of attention from fact-checking systems. And the influence typically depends on two key factors: popularity and type. However, due to the lack of social hotspot indicators and category classification, existing benchmarks evaluate fact-checking systems equally across all claims, failing to assess the system's performance in handling high-influence events. Moreover, most existing datasets focus on English-language contexts, and the lack of Chinese benchmarks hinders the comprehensive development of AI in fact-checking.

We define the fact-checking system's Hotspot Perception Ability (HPA) as its effectiveness in handling high-influence events and propose TrendFact, the first benchmark capable of evaluating HPA and covering all fact-checking tasks. TrendFact focuses on the Chinese language, contains 7,643 samples with an evidence library of 366,634 pieces, and supports all fact-checking tasks, including fact verification, evidence retrieval, and explanation generation. Each sample includes 10 attributes, as shown in Figure 1. Among them, the influence score is derived from the risk score and hotspot indicators. Furthermore, TrendFact evaluates HPA through our proposed metric, HCPI (Hotspot Claim Perception Index), and introduces the ECS (Explanation Consistency Score) metric to fill the gap in evaluating explanation generation consistency.

The construction of the TrendFact benchmark consists of four components: (1) Data Collection, (2) Data Augmentation, (3) Evidence Library Construction, and (4) ECS and HCPI Metrics Definition. First, we gather large-scale real-world fact entries from sources like Weibo, Baidu, etc., along with the existing CHEF dataset as the initial data. Then, we apply two separate augmentation processes to the two data sources. For the construction of evidence library, we combine gold evidence from the dataset with evidence retrieved through our defined process. Finally, we provide formal definitions for the ECS and HCPI metrics.

Furthermore, to address the performance degradation of RLMs caused by the predefined fixed retrieval in traditional retrieve augmentation generation (RAG), as discussed in previous work (Li et al., 2025), we propose FactISR, a reasoning framework that enhances RLMs' performance and HPA on fact-checking tasks by combining dynamic evidence augmentation with influence score-based iterative self-reflection. We evaluate TrendFact on advanced LLMs, RLMs and fact-checking methods. The experiment results show that most tested methods exhibit limited performance, and our FactISR achieves performance improvements with improved computational efficiency.

In summary, our contributions are as follows:

- We first define the fact-checking system's Hotspot Perception Ability (HPA) as its effectiveness in handling high-influence events, and introduce the first Chinese-language benchmark that provides comprehensive evaluations on evidence retrieval, fact verification, explanation generation and HPA, along with two new metrics: ECS for evaluating explanation generation consistency and HCPI for assessing HPA.
- We propose FactISR, a reasoning framework that enhances RLMs on fact-checking tasks by combining dynamic evidence augmentation with influence score-based iterative self-reflection.
- We conduct extensive experiments that reveal limitations in the capabilities of LLMs, RLMs, and fact-checking methods on TrendFact. Furthermore, the results show that FactISR can enhance the performance of RLMs.

2 Related Work

Fact-checking Benchmarks Existing fact-checking benchmarks are generally divided into two categories: one based on Wikipedia data, such as Hover (Jiang et al., 2020) and FEVER (Thorne et al., 2018), and another using knowledge bases from fact-checking websites, such as CLAIMDECOMP (Chen et al., 2022) and QUANTEMP (Venktesh et al., 2024) (Table 1). These benchmarks primarily focus on fact verification and evidence retrieval tasks, often neglecting explanation generation. With the growing role of LLMs and RLMs in generating explanations, evaluating their reliability is crucial. Moreover, it is also a pressing issue in the field to evaluate the HPA of fact-checking systems (Solovev and Pröllochs, 2022; Sehat et al., 2024), which refers

Dataset	#Claims	Source	Language	Explanation Contains	HPA	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. WP refers to Wikipedia, FCS refers to fact-checking websites, FB refers to FreeBase, and TP refers to Trending Platforms. By 'HPA', we refer to whether the dataset supports the evaluation of a fact-checking system's hotspot perception ability.

to the system's greater effort in addressing high influence events. Few benchmarks (Gupta and Srikumar, 2021; Lin et al., 2024; Hu et al., 2022) have addressed datasets in non-English languages. Therefore, developing a Chinese benchmark like TrendFact, which evaluates both explanation generation and hotspot perception, is essential for comprehensive fact-checking evaluation.

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 and RLMs, methods like PROGRAMFC (Pan et al., 2023) generate 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) generates counter-misinformation responses to correct false claims. Additionally, many methods overlook the HPA, which is essential for improving fact-checking systems' ability to solve highly influential claims. Incorporating HPA, along with the interplay between verification and explanation, can enhance transparency, trust, and the solving capability of high influence claims in the real world.

3 HotSpot Perception Ability (HPA)

With the explosive growth of information and the accelerated spread on social media, fact-checking systems face significant challenges and need to account for the variations in social influence across events (Solovev and Pröllochs, 2022; Sehat et al., 2024; Damasceno et al., 2024). For high-influence events, such as misinformation in politics and public health, the system should prioritize these events and allocate more resources. To this end, we introduce the Hotspot Perception Ability (HPA) to measure a fact-checking system's capacity under such conditions, which we formally define as:

$$\text{HPA} = f(\text{C}, \text{HI}, \text{R}) \quad (1)$$

Where C means event category, HI refers to the Hotspot Indicators of an event (e.g., popularity, discussion volume), and R represents the computational resources. By enhancing HPA, the system can dynamically adjust its resource allocation, improving the effectiveness and efficiency of handling events with varying influence.

4 TrendFact Construction

In this section, we detail the overall construction process of TrendFact, as illustrated in Figure 2. We begin by collecting claims from two different sources (Section 4.2) and applying two separate augmentation processes to refine the collected claims (Section 4.3). Then, we construct an evidence library (Section 4.4). Finally, to enable

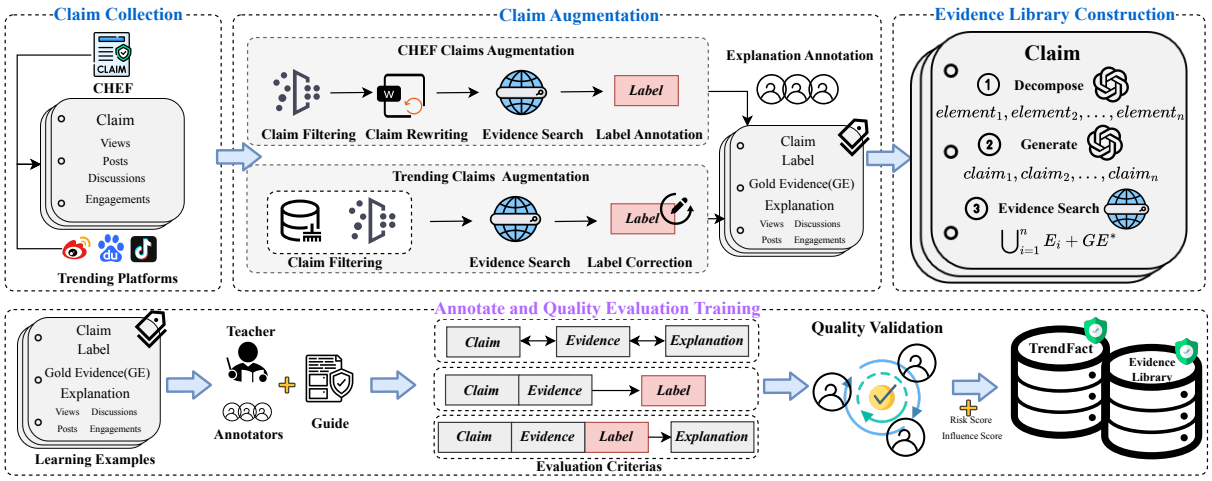

Figure 2: The overall construction process of TrendFact includes claim collection, filtering, augmentation, evidence library construction, and a multi-stage sample review process.

TrendFact to evaluate both the HPA and the explanation generation consistency of fact-checking systems, we introduce two new metrics: ECS and HCPI (Section 4.5).

4.1 Data Attributes

TrendFact covers five domains: public health, science, society, politics, and culture. As shown in Figure 1, each sample is annotated with attributes including claim, label, gold evidence, textual explanation, risk score, influence score, and four hotspot indicators. Based on the number of gold evidence items, samples are categorized into single-evidence and multi-evidence, with single-evidence samples accounting for 85%. For more detailed definitions of hotspot indicators, influence scores, and other attributes, see Appendix A.

4.2 Claim Collection

In contrast to previous datasets that primarily relied on sources with limited timeliness and popularity, such as Wikipedia and fact-checking websites, TrendFact collects claims from two complementary sources. **Source One: Trending Platforms.** We collect claims from Weibo, DouYin, and Baidu, which provide large-scale, dynamic factual statements accompanied by real-time hotspot indicators (e.g., view counts and discussion volumes). These features are crucial for evaluating the HPA of fact-checking systems. In total, we obtain approximately 500,000 claims from these platforms between 2020 and 2024. **Source Two: Existing Datasets.** We further collect claims from existing fact-checking datasets, which broaden verification scenarios beyond trending platforms. Although such datasets lack hotspot indicators, they still offer valuable resources. In particular, we incorporate claims from the CHEF dataset, which aggregates data from multiple fact-checking websites, as a representative supplement.

4.3 Data Augmentation

The initial claims collected from trending platforms and existing datasets exhibit distinct limitations. Claims from trending platforms often lack the key attributes required for fact-checking (e.g., labels, evidence, and explanations) and contain a large proportion of meaningless noise. In contrast, claims from the existing dataset CHEF, while annotated with labels and evidence, frequently include judgmental words (e.g., "... is a rumor"), noisy samples, and lack textual explanations. To address these issues, we design tailored augmentation processes for the two claim sources.

Trending Claims. We first apply an LLM voting mechanism to filter out approximately 90% of low-quality or trivial claims (the filtering principle is illustrated in Appendix Table 8). Then, our annotation team removes sensitive or overlapping content, resulting in a set of 6,512 claims. Since the remaining claims often lack the essential factual elements and cannot be used directly as fact-checking samples, annotators follow a rewriting guideline to produce qualified claims. This rewriting process adheres to five fact-checking claim rewriting factors, including: **Temporal Anchoring:** Adding or specifying temporal references. **Data Granularity:** Breaking down aggregated data into verifiable units. **Ambiguity Resolution:** Replacing fuzzy

qualifiers with deterministic statements. **Comparative Standard:** Introducing quantifiable comparison objects and proportions. **Domain Knowledge:** Supplementing with industry-specific parameters or mechanisms. More details are in Appendix C.

For each rewritten claim, annotators then retrieve one piece of gold evidence and provide a detailed manual explanation. More examples of filtering and rewriting are shown in Appendix Figure 7 and Figure 8, respectively.

CHEF Claims. We first remove sensitive or factually inaccurate samples and then use the same LLM voting mechanism to select claims that present significant verification challenges. Each selected claim is further annotated with a detailed explanation by our annotation team that shares the same standard applied to claims. This process results in 1,131 enhanced claims.

After augmentation, we merge the two sets of claims to construct TrendFact, which consists of 7,643 samples. To ensure the quality of the augmented data, annotators underwent rigorous training, including understanding and applying the filtering mechanisms, rewriting guidelines, and reviewing the claim annotations. The training process also covered examples of proper claim rewriting, emphasizing the importance of maintaining semantic integrity while transforming the claim into a verifiable, fact-checkable statement. As shown in Figure 2, all rewritten and annotated outputs are validated through a three-level standardized human evaluation criteria, ensuring consistency and reliability across the dataset (see Appendix D for details).

4.4 Evidence Library Construction

As illustrated in Figure 2, the evidence library construction of the TrendFact follows a three-stage process: decomposition, generation, and retrieval. Specifically, we first employ an LLM to extract three or more key elements from each claim. We generate new claims for elements and retrieve supporting evidence from websites, which is then combined with the gold evidence from the initial TrendFact dataset to form a comprehensive evidence library containing 366,634 entries. More detailed descriptions are provided in Appendix E.

4.5 ECS and HCPI Metrics Definition

To bridge the gap of evaluating explanation generation consistency and HPA of fact-checking systems, we introduce ECS and HCPI, respectively.

Category	Label Acc	Explain Cons	Score
F-D	F	Full Discrepancy	0.2
F-C	F	Consistency	0.4
T-CD	T	Content Divergence	0.6
T-PC	T	Partial Consistency	0.8
T-FC	T	Full Consistency	1.0

Table 2: Definition of ECS. Label Acc indicates the accuracy of the fact verification, and Explain Cons represents the consistency of the explanation generated by the system and the gold one that evaluated by LLM .

ECS As shown in Table 2, we define five categories to describe explanation consistency based on the consistency between the gold explanation and the generated explanation with system's verification accuracy: (1) **F-D**: Incorrect verification with full discrepancy; (2) **F-C**: Incorrect verification but consistent explanation; (3) **T-CD**: Correct verification with conflict explanation; (4) **T-PC**: Correct verification but partial consistency; (5) **T-FC**: Correct verification with full consistency. Higher explanation quality corresponds to higher scores when the system's verification is correct.

HCPI The sample attributes of TrendFact align well with the key requirements for evaluating a fact-checking system's Hotspot Perception Ability (HPA), as discussed in Section 1. Based on this connection, we introduce the Hotspot Claim Perception Index (HCPI) to quantitatively assess HPA by considering the social influence contributions of different claim types: Support (SUP), Refute (REF), and Not Enough Information (NEI). Overall, we define HCPI as follows:

$$\text{HCPI} = \frac{S_{\text{SUP}} + S_{\text{REF}} + S_{\text{NEF}}}{\sum\limits_{i \in \text{All Claims}} \text{s}_i} \tag{2}$$

Here, the three types of claim are weighted by different penalty factors in their social influence score calculations (the detailed formulas are in Appendix F). s_i denotes the social influence score of the i-th sample, calculated from the weighted contributions of four hotspot indicators (views, discussions, engagements, and posts), combined with the risk score evaluated by GPT4.1, defined as:

$$\text{s}_i = r_i \cdot \sum_{x \in \{v_i, d_i, e_i, p_i\}} w_x \cdot \log(1 + x) \tag{3}$$

In summary, higher HCPI values correspond to stronger HPA performance of the system.

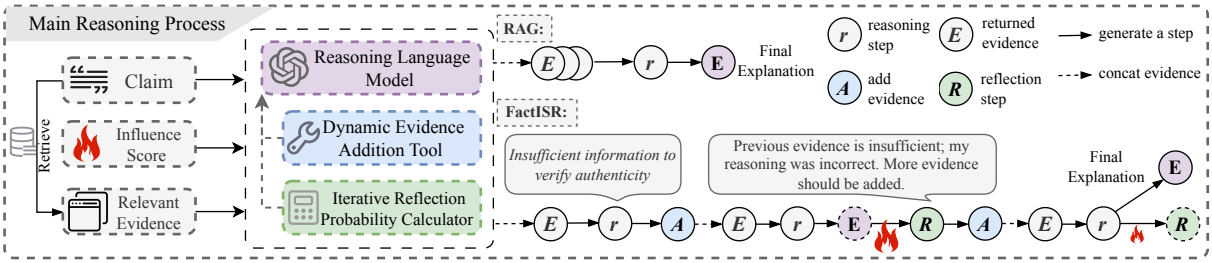

Figure 3: Overview of FactISR. The bottom-right section shows the reasoning process of FactISR, where "flame" represents the reflection probability calculated based on influence score, which continuously decays with the number of iterative reflections. The reasoning process terminates when the maximum number of reflections is exceeded or no reflection step is sampled.

5 FactISR

In this section, we detail our proposed method, FactISR, a reasoning framework designed to enhance the fact-checking capabilities and HPA of RLMs. As shown in Figure 3, it consists of two key components: dynamic evidence augmentation (DEA) and iterative self-reflection (ISR). Unlike traditional RAG, which loads all evidence at once for reasoning, FactISR dynamically calls upon DEA to supplement evidence throughout the reasoning process. And based on the claim's influence score and the interim reasoning results, ISR is selectively activated by RLM to further refine the process and achieve more accurate outcomes.

Dynamic Evidence Augmentation Traditional RAG process loads all evidence at once, which can lead to issues such as insufficient, redundant, or irrelevant evidence, hindering accurate judgment. To address these, FactISR's DEA module dynamically retrieves and adds relevant evidence during the reasoning process based on ongoing analysis of RLMs. This continuous addition of evidence guarantees a more refined evidence and produces better fact-checking performance.

Iterative Self-Reflection Traditional RAG evaluates all claims equally and lacks the capability to adjust its reasoning budget based on a claim's popularity. This limitation can lead to redundant inference for low-influence claims and insufficient reasoning for high-influence ones, ultimately detracting from HPA performance. To address these, we introduce the iterative self-reflection module, which enables RLMs to iteratively reflect according to a claim's popularity.

Specifically, if a claim has a high influence score, the probability of reflection reasoning increases. Moreover, as the reasoning process advances, the probability gradually decreases, effectively preventing excessive reflection. The formal definition of the reflection probability is as follows:

$$P_i^n = (k \cdot \frac{s_i - s_{\min}}{s_{\max} - s_{\min} + \epsilon}) \cdot \gamma^{n-1} \quad (4)$$

Where P_i^n represents the probability of the n-th reflection for the i-th sample, while $s_{\min}$ and $s_{\max}$ denote the minimum and maximum scores among all claims, respectively. A small constant ϵ is included to prevent division by zero, typically set to 1×10^{-8}. The hyperparameter $k \in [0, 1]$ is crucial for flexibly controlling the overall reflection intensity across samples, with a default value of 1, and $\gamma \in [0, 1]$ is the decay factor. where the superscript n in γ^n indicates exponentiation.

6 Experiment

6.1 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 explanation generation task, in addition to ECS, we also employ BLEU-4, ROUGE-(1, 2, L), and BERTScore. For assessing the HPA of fact-checking systems, we employ HCPI.

Baselines In this work, we choose the following types of 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 et al., 2025), and DeepSeek-R1-0528 (Guo et al., 2025). For LLMs, we choose GPT-4.1 (Hurst et al., 2024), DeepSeek-v3 (Liu et al., 2024), and Qwen2.5-72B-Instruct (qwe, 2024), Qwen3-32B (No Think). For fact-checking methods, we select PROGRAMFC (Pan et al., 2023) and CLAIMDECOMP (Chen et al., 2022).

Methods	R@1	R@2	R@3	R@5
BM25 -w/o *date*	12.08	20.46	25.98	33.70
BM25	12.99	20.76	26.65	34.91
text-emb-ada-002	12.10	21.44	26.88	35.42
bge-m3(dense)	**14.10**	**22.76**	**29.63**	**39.97**

Table 3: Experimental Results on Evidence Retrieval.

Methods	F1	P	R	Acc
PROGRAM-FC	40.46	41.66	43.30	43.17
CLAIMDECOMP	42.53	44.18	44.90	45.28
QwQ-32B-Preview	49.52	52.32	53.48	55.29
Qwen2.5-72B-instruct	47.10	53.51	52.51	55.96
Qwen3-32B(*No think*)	50.06	53.00	53.92	58.84
DeepSeek-V3-0324	52.10	55.19	55.94	60.67
GPT-4.1	52.45	56.23	55.88	61.29
Qwen3-32B(*Think*)	58.14	58.05	59.82	66.09
QwQ-32B	58.58	58.89	60.09	68.61
DeepSeek-R1-0528	58.89	59.15	60.58	68.44
FactISR(*Qwen3-32B*)	58.68	58.55	60.84	67.49
FactISR(*QwQ-32B*)	**61.17**	**61.04**	**63.37**	**69.70**

Table 4: Comparison of FactISR with other baselines on fact verification task.

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 The detailed experiment settings are provided in the Appendix J.

6.2 Main Results

Evidence Retrieval Results We evaluate the selected retrieval methods based on their ability to retrieve the target gold evidence from the TrendFact evidence library. As shown in Table 3, the best performing method, bge-m3, achieves a relatively low performance with an R@5 less than 40%. This suggests that our unique evidence library construction effectively gathers challenging evidence that distinguishes original gold evidence, thereby increasing the difficulty.

Fact Verification Results We conduct a comprehensive evaluation of selected methods on the verification task of TrendFact, with the results presented in Table 4. The key findings are as follows: First, traditional fact-checking methods perform the worst, with all verification scores falling below 50%. Second, both LLMs and RLMs perform better than 50%. Notably, RLMs outperform LLMs, as they are better equipped to handle the complex reasoning required by many TrendFact samples. However, even the best-performing DeepSeek-R1-0528 fails to achieve an overall verification score above 60%, highlighting the significant challenge posed by the high-quality and reasoning-intensive nature of TrendFact. Moreover, FactISR contributes to improved verification performance for RLMs. For instance, it enhances the performance of the RLM QwQ-32B, enabling it to achieve an overall verification F1 score of 61, surpassing DeepSeek-R1-0528.

Explanation Generation Results We evaluate both LLMs and RLMs on the explanation generation task in TrendFact, with the results presented in Table 5. Our findings are as follows: First, RLMs generally produce lower-quality explanations compared to LLMs. We attribute this to the design of RLMs, which prioritizes concise key information and accurate conclusions over elaborate output, resulting in less detailed explanations. Second, RLMs outperform LLMs in explanation consistency, as measured by ECS. This is expected, as ECS reflects the alignment between explanations and the model's internal reasoning process. Furthermore, FactISR enhances both the explanation quality and consistency of RLMs. For example, FactISR improves the performance of QwQ-32B by 1-3 percentage points across all explanation metrics, enabling it to approach the performance of LLMs in explanation quality.

HPA Assessment Results Since the verification results affect the calculation of HCPI, the RLM naturally exhibits a higher HCPI compared to the LLM. However, because the RLM lacks the capability to adjust the reasoning budget according to the claim's influence, it may result in redundant inference for low-influence claims and insufficient reasoning for high-influence claims, adversely affecting its HPA performance. In contrast, FactISR effectively addresses these issues by leveraging the influence score-based ISR module combined with DEA, and improves the RLM's HCPI score by up to nearly 4%.

6.3 Ablation Study

Component Effectiveness We conduct experiments by individually removing one of the two modules from the fully integrated QwQ-32B and Qwen3-32B. We evaluate comprehensive performance using accuracy, ECS, and HCPI. The results

Methods	HCPI	ECS	BLEU-4	BERTScore	ROUGE-1	ROUGE-2	ROUGE-L
QwQ-32B-Preview	0.4820	0.7843	0.1573	0.7525	0.4699	0.2781	0.4048
Qwen2.5-72B-instruct	0.5061	0.7207	**0.2632**	0.8015	0.5637	**0.3712**	0.5080
Qwen3-32B(*No think*)	0.5001	0.7574	0.2491	**0.8049**	**0.5649**	0.3658	**0.5106**
DeepSeek-V3-0324	0.5427	0.7674	0.2360	0.7918	0.5423	0.3427	0.4887
GPT-4.1	0.5488	0.7726	0.2214	0.7735	0.5380	0.3341	0.4736
Qwen3-32B(*Think*)	0.5575	0.8332	0.2181	0.7836	0.5188	0.3138	0.4603
QwQ-32B	0.5621	0.8531	0.1940	0.7712	0.4896	0.2864	0.4247
DeepSeek-R1-0528	0.5681	0.8399	0.2041	0.7764	0.5016	0.2954	0.4383
FactISR(*Qwen3-32B*)	0.5926	0.8426	0.2281	0.7912	0.5365	0.3334	0.4833
FactISR(*QwQ-32B*)	**0.6010**	**0.8610**	0.2175	0.7835	0.5156	0.3107	0.4592

Table 5: Comparison of FactISR with Other Baselines on Explanation Generation.

Methods	F1	ECS	HCPI
FactISR(*QwQ-32B*)	**61.17**	**0.8610**	**0.6010**
- w/o *DEA*	59.79	0.8501	0.5833
- w/o *ISR*	58.88	0.8388	0.5675
FactISR(*Qwen3-32B*)	**58.68**	**0.8426**	**0.5926**
- w/o *DEA*	58.46	0.8372	0.5785
- w/o *ISR*	58.30	0.8390	0.5650

Table 6: Ablation Study for Evaluating Each Component of method FactISR.

are shown in Table 6. It demonstrates that removing any single module leads to performance degradation across all tasks, confirming the effectiveness of all components in FactISR.

DEA Efficiency We conduct experiments to analyze the efficiency gains introduced by the DEA module. As shown in Table 7, it contributes to maintaining accuracy while achieving substantial gains in reasoning efficiency, with maximum reductions of 30% and 50% in reasoning time and length, respectively. These results indicate that DEA effectively mitigates the reasoning inefficiencies caused by lengthy evidence.

Evaluation with Limited Resources We conduct experiments to evaluate the influence of FactISR on the HPA performance and fact-checking efficiency of RLMs under resource constraints, as shown in Figure 4. We adjust the number of input evidence for RAG and the reflection intensity hyperparameter k for FactISR. It can be seen that FactISR achieves higher HPA with the same amount of evidence and significantly reduces the average inference time required to reach the same HPA, thereby greatly improving verification efficiency.

Methods	F1	Time	Length
QwQ-32B	**58.89**	0.5960	2664
+ *DEA*	58.88	**0.4594** (↓22.92%)	**1442** (↓47.31%)
Qwen3-32B	58.14	0.5515	2664
+ *DEA*	**58.30**	**0.4380** (↓30.17%)	**1316** (↓50.62%)

Table 7: Impact of DEA on Per-Sample Generation Time and Input Evidence Length.

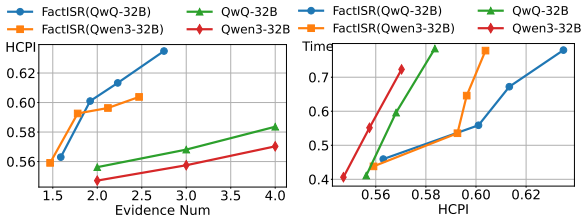

Figure 4: Performance comparison between RAG and FactISR under resource constraints.

7 Conclusion

In this paper, we introduce TrendFact, the first hotspot perception fact-checking benchmark for the evaluation of fact verification, evidence retrieval, and explanation generation tasks. It comprises 7,643 challenging samples and an evidence library containing 366,634 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 performance.

8 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.

9 Ethics and Compliance

The claims in TrendFact are derived from trending platforms and existing datasets, with evidence sourced from publicly content on the Internet, and its construction does not violate relevant platform rules or legal regulations (Calzada, 2022; Chen and Sun, 2021), making it suitable for academic research. Specifically, the claims collected from trending platforms are taken exclusively from public pages that do not require login. The collection process strictly follows the platform's open-access agreements (e.g., Weibo Open Platform Agreement: https://open.weibo.com), and only public data is retrieved, without involving any non-public interfaces, private user content, or data requiring authorization. During it, we only collected information such as trending titles, dates, and urls, which fall within the lowest-risk category, and do not include sensitive data such as user nicknames or user IDs. Date information is retained only to the level of "year–month–day" rather than exact timestamps; additionally, we applied rate limiting (≤ 1 request per second) to avoid imposing load on the servers.

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*.

Igor Calzada. 2022. Citizens' data privacy in china: The state of the art of the personal information protection law (pipl). *Smart Cities*, 5(3):1129–1150.

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*.

Jihong Chen and Jiabin Sun. 2021. Understanding the chinese data security law. *International Cybersecurity Law Review*, 2(2):209–221.

Lucas P Damasceno, Egzona Rexhepi, Allison Shafer, Ian Whitehouse, Nathalie Japkowicz, Charles C Cavalcante, Roberto Corizzo, and Zois Boukouvalas. 2024. Exploiting sparsity and statistical dependence in multivariate data fusion: an application to misinformation detection for high-impact events. *Machine Learning*, 113(4):2183–2205.

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*.

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

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.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

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*.

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.

Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X Zhang. 2024. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proceedings of the ACM on human-computer interaction*, 8(CSCW1):1–36.

Kirill Solovev and Nicolas Pröllochs. 2022. Moral emotions shape the virality of covid-19 misinformation on social media. In *Proceedings of the ACM web conference 2022*, pages 3706–3717.

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.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

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 indicators of the 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. Figure 5 shows the data distribution of TrendFact samples, including labels, gold evidence count, and domains.

B Data Cleaning and Hard Example Selection

Figure 6 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

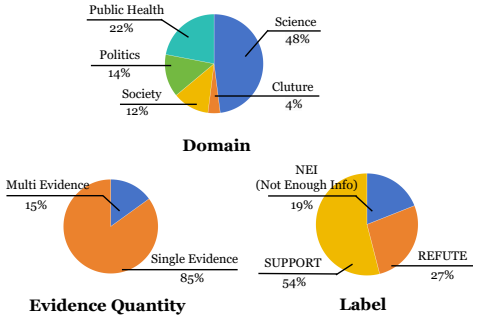

Figure 5: Overview of the data distribution, including labels, gold evidence count, and domains.

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 8 and Figure 7 respectively present examples of stage-specific prompts and the sample filtering workflow.

C Fact-Checking Claim Rewriting Factor

As shown in Table 9, 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 8.

D Annotate Evaluation Criteria

Our human evaluation criteria, as illustrated in Figure 2, 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

First, we employ GPT-4.1 to extract multiple key-element phrases from the original claims and to generate rewritten statements based on the original claims. Figure 10 showcases 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 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 HCPI

Equation 2 provides a detailed definition of the HCPI scoring mechanism. The descriptions of the three sample types are as follows: **Support (SUP)**: Samples originally labeled as support; **Refute (REF)**: Samples originally labeled as refute; **Not Enough Information (NEI)**: Samples where there is insufficient evidence to determine the authenticity of the claim.

And, the social influence score calculations for them are calculated as follows:

$$S_{\text{SUP}} = \sum_{i \in \text{SUP}} \begin{cases} \mathrm{s}_i \cdot \text{ECS}(i), & \hat{y}_i = \text{SUP} \\ -2 \cdot \mathrm{s}_i, & \hat{y}_i = \text{REF} \end{cases} \tag{5}$$

$$S_{\text{REF}} = \sum_{j \in \text{REF}} \mathrm{s}_j \cdot \text{ECS}(j) \cdot \delta(\hat{y}_j = \text{REF}) \tag{6}$$

$$S_{\text{NEI}} = \begin{cases} -\mathrm{s}_k, & \hat{y}_k = \text{REF} \\ \mathrm{s}_k \cdot \text{ECS}(k), & \hat{y}_k = \text{NEI} \end{cases} \tag{7}$$

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: AmessagetoourcommunityReadourCOVID-19Responsehere.©Copyright2021InteracInteracandInterace-Transferareregisteredtrade-marksandtheInteraclogo,InLifeandInLifedesignaretrademarksofInteracCorp.ThesecookiesareessentialforourSitestooperateproperly.Essentialcookiesensurebasicfunctionalitiesandsecurityfeaturesofthe Sites.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 6: Examples of CHEF Dataset Cleaning.

For the SUP (Support) samples, if they are correctly predicted, their social influence score is calculated as the product of Inf(i) and ECS(i). However, if the prediction is incorrect (i.e., predicted as REF), a higher penalty is applied, represented by -2 * Inf(i). This ensures that misclassifying a positive sample incurs a larger penalty. For the REF (Refute) samples, the social influence score is computed using the product of Inf(j), ECS(j), and the delta function $\delta(\hat{y}_j = \text{REF})$. This term captures the influence of REF samples, considering their relevance and correctness in prediction. For the NEI (Not Enough Information) samples, if misclassified as REF, the score is penalized as -Inf(k). If correctly classified as NEI, the score is calculated as Inf(k) * ECS(k), ensuring that the contribution to the score is aligned with the quality of the explanatory content.

G Prompt of FactISR

Figure 11 shows the prompt of FactISR.

H Example of FactISR

Figure 13 illustrates an example of FactISR. Without ISR, the model directly outputs a conclusion of insufficient evidence and prematurely ends the reasoning process. Our reflection mechanism 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 0.3573, which is nearly 0.1 points higher than when using retrieval-based evidence.

J Details of the Experimental Settings

We conduct experiments on PyTorch[1] and $2 \times A100$ GPUs. The evaluation for ESC was conducted using GPT-4.1, while BERTScore evaluations are conducted on *chinese-bert-Base*[2]. The small constant ϵ to prevent division by zero, the hyperparameter k controlling the reflection strength, and the decay factor γ are set to 1×10^{-8}, 1 and 0.5,

[1] https://pytorch.org/
[2] https://huggingface.co/google-bert/bert-base-chinese

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

respectively. The maximum number of reflections is set to 3. 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.

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

Figure 9: Evidence Library Construction Process.

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 8: 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 9: Fact-Checking Claim Rewriting Factor

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-4.1	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.

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-4.1	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.

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 10: Claim Rewriting Prompt for Evidence Retrieval.

Prompt of FactISR

Task Description

You are a fact-checking expert with retrieval capabilities. You need to determine the veracity of a given statement based on evidence and provide a textual explanation.

The labels for statements are limited to three types: **true**, **false**, or **insufficient evidence**.
The system will provide initial evidence. If the existing evidence is insufficient to determine the veracity, you must use retrieval tools to gather additional evidence.

Tools

I will provide you with retrieval tools to add additional evidence.

Output Format

Your output should follow one of the two formats below:
<tool_call>
Add additional evidence
</tool_call>
or
<explanation>
Generate a textual explanation for the confirmed answer to present to the user.
</explanation>

Notes

1. You must strictly adhere to the two formats above.
2. Do not mention in your response that you need to call search tools!
3. If the additional evidence returned by the system is still insufficient to determine the veracity of the statement, you should continue using the tools to add evidence rather than responding directly. You can call the tools multiple times until the system reaches the maximum call limit.
4. Some statements have a broader impact and wider dissemination. Users may ask you to reconsider these statements. In such cases, prioritize responding to the user's request. If you find no issues with the previous reasoning logic, simply repeat the previous textual explanation in the correct format, i.e., <explanation>...</explanation>. If you identify errors in the previous reasoning, correct them promptly and proceed with the next step (continue calling tools to gather more information or reassess the veracity based on existing information and generate a new explanation in the correct format).
5. When making time-related judgments, if the publication date is empty, use the actual date mentioned in the evidence. Otherwise, assume the current evidence aligns with the statement in terms of time.
6. Never conclude "insufficient evidence" without collecting adequate evidence first.
7. Labeling an ambiguous statement as false (a rumor) can have more severe consequences unless the evidence contains sufficient details to justify labeling it as false.
8. Keep the final explanation concise, ideally within 150 words, and no more than 300 words!

Fianl Explanation Output Format

1. **True**:
   "<explanation>The evidence indicates [...] Therefore, the claim is **substantiated**.</explanation>"
2. **False**:
   "<explanation>The evidence indicates [...] Therefore, the claim is **refuted**.</explanation>"
3. **Insufficient Evidence**:
   "<explanation>The evidence indicates [...] Therefore, the claim **cannot be verified** due to insufficient evidence.</explanation>"

Figure 11: 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.***

</think>

'tool_calls': [{'type': 'function','function': {'name': 'AddEvidence', 'arguments': {'num': 1}}}]

{'role': 'tool', 'name': 'AddEvidence', 'content': '.....'}

<think>

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.***

</think>

'tool_calls': [{'type': 'function','function': {'name': 'AddEvidence', 'arguments': {'num': 2}}}]

{'role': 'tool', 'name': 'AddEvidence', 'content': '.....'}

<think>

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 12: Example of Dynamic Evidence Addition.

FactISR Example

Claim: *On May 2, 2022, the price of electricity in France surged to four times its historical lowest price.*

<think>
Okay, I need to determine whether the statement "*On May 2, 2022, the price of electricity in France surged to four times its historical lowest price.*" is true. First, let's look at the initial evidence, which mentions that French electricity prices surged by 400%, and that May 2 became a trending topic. The evidence states that the price rose to four times that of the same period last year, but the statement refers to "four times the historical lowest price." There might be a crucial difference here. Are the historical lowest price and the price from the same period last year the same? If the French electricity price in the same period last year was not the historical lowest, then the statement might be inaccurate. It is necessary to confirm whether the electricity price on May 2, 2022, was indeed four times the historical lowest price. This may require searching for historical data on French electricity prices, especially the lowest price in the past ten years, as well as the actual price in May 2022. If the evidence does not explicitly mention the historical lowest price, further research may be needed. However, the current evidence only mentions that the price rose to four times that of the same period last year, without mentioning the historical lowest price, so the statement may be inaccurate. It is necessary to use tools to obtain more information.

</think>

{'tool_calls': [{'type': 'function','function': {'name': 'AddEvidence', 'arguments': {'num': 1}}}]}

{'role': 'tool', 'name': 'AddEvidence', 'content': '.....'}

<think>
Okay, now I need to analyze the user's statement and the provided evidence again. The user's statement is: "On May 2, 2022, French electricity prices soared to four times the historical lowest price." Both the initial and supplementary evidence mention that French electricity prices surged by 400% on May 2, 2022, which means they rose to four times the price of the same period last year. However, the issue is that the statement refers to "four times the historical lowest price," while the evidence does not mention the historical lowest price, only making a comparison with the same period last year. Therefore, it is necessary to confirm whether the electricity price during the same period last year was indeed the historical lowest. If the price during the same period last year was already the historical lowest, then the statement is accurate; otherwise, the statement may be inaccurate. Since the existing evidence does not provide historical price data, further investigation is needed to find the historical lowest point of French electricity prices, but the user may not be able to provide more evidence. Therefore, based on the current information, it is not possible to determine whether the price was four times the historical lowest; we can only conclude that the evidence is insufficient
</think>
<explanation>
The evidence shows that on May 2, 2022, French electricity prices rose to four times the level of the same period last year, but did not mention the specific value of the "historical lowest price." Therefore, the evidence is insufficient to verify the accuracy of the current statement.
<explanation>

Reflection

<think>
I should reflect on my previous reasoning to ensure the accuracy of my judgment. It is correct that the evidence did not mention the specific value of the "historical lowest price." However, I made the mistake of drawing a conclusion prematurely after adding only one piece of evidence. In fact, when evidence is insufficient, I should continue to add more evidence until there is clear information to support my judgment or until the maximum number of additions is reached. Therefore, regarding the current statement, I should continue to use tools to obtain more information.
</think>

{'tool_calls': [{'type': 'function','function': {'name': 'AddEvidence', 'arguments': {'num': 1}}}]}

{'role': 'tool', 'name': 'AddEvidence', 'content': '.....'}

<think>
Okay, based on the newly supplemented information, I have obtained the historical lowest price of French electricity. On May 2, 2022, the French electricity price was much more than four times the historical lowest price. Therefore, this statement can be judged as false.
</think>
<explanation>
The evidence shows that on May 2, 2022, French electricity prices surged to four times the level of the same period last year, not four times the historical lowest price. Therefore, this statement is incorrect.
<explanation>

Figure 13: An Example of FactISR.