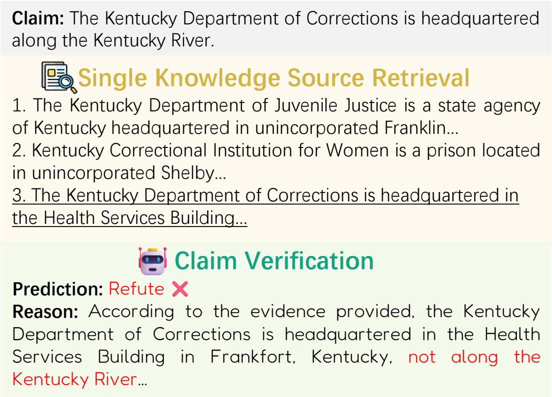
IMRRF: Integrating Multi-Source Retrieval and Redundancy Filtering for LLM-based Fake News Detection

Dayang Li[1,2], Fanxiao Li[2,3], BingBing Song[1,2], Li Tang[1,2], and Wei Zhou[1,2]*
[1]National Pilot School of Software, Yunnan University, Yunnan, China
[2] Engineering Research Center of Cyberspace, Yunnan University, Yunnan, China
[3]School of Information Science and Engineering, Yunnan University, Yunnan, China
lidayang1@stu.ynu.edu.cn,zwei@ynu.edu.cn

Abstract

The widespread use of social networks has significantly accelerated the dissemination of information but has also facilitated the rapid spread of fake news, leading to various negative consequences. Recently, with the emergence of large language models (LLMs), researchers have focused on leveraging LLMs for automated fake news detection. Unfortunately, many issues remain to be addressed. First, the evidence retrieved to verify given fake news is often insufficient, limiting the performance of LLMs when reasoning directly from this evidence. Additionally, the retrieved evidence frequently contains substantial redundant information, which can interfere with the LLMs' judgment. To address these limitations, we propose a Multiple Knowledge Sources Retrieval and LLM World Knowledge Conversion framework, which enriches the evidence available for claim verification. We also introduce a Redundant Information Filtering strategy, which minimizes the influence of irrelevant information on the LLM reasoning process. Extensive experiments conducted on two challenging fact-checking datasets demonstrate that our proposed method outperforms state-of-the-art fact-checking baselines. Our code is available at https://github.com/quark233/IMRRF.



Claim: The Kentucky Department of Corrections is headquartered along the Kentucky River.

Single Knowledge Source Retrieval

1. The Kentucky Department of Juvenile Justice is a state agency of Kentucky headquartered in unincorporated Franklin...
2. Kentucky Correctional Institution for Women is a prison located in unincorporated Shelby...
3. The Kentucky Department of Corrections is headquartered in the Health Services Building...

Claim Verification

Prediction: Refute ✗
Reason: According to the evidence provided, the Kentucky Department of Corrections is headquartered in the Health Services Building in Frankfort, Kentucky, not along the Kentucky River...

Figure 1: Evidence retrieved from a single knowledge source may be incomplete, which can affect the validation process.

1 Introduction

With the rapid development of social networks, information exchange has become more convenient. However, the risk of fake news also raises concerns. Fake news is often misleading, deceptive, and sometimes malicious. It not only misleads the public and causes confusion but also significantly influences public opinion. Previous fake news detection methods primarily relied on manual checking, which is time-consuming. Therefore, developing effective automated fake news detection methods is crucial for curbing the spread and ensuring the authenticity of information.

Early efforts for automatic detection of fake news (Ma et al., 2016; Yu et al., 2017; Cheng et al., 2020; Trueman et al., 2021; Jang et al., 2022) mainly focused on analyzing news content and using deep learning methods to determine the authenticity of the news. Although such methods are effective, they make predictions only rely on the internal knowledge of the used models. Alhindi et al. (2018) demonstrated that the incorporation of external news-related evidence could enhance detection performance and provide factual support. Consequently, fact-checking based methods (Thorne and Vlachos, 2018; Nakov et al., 2021) have been the mainstream for fake news detection. Despite significant progress that has been made, training such detection models require massive labeled data, which is time-consuming and labor-intensive. In addition, such methods (Cui et al., 2019) produce only one classification label without human-readable explanations, which makes it difficult for fact-checkers to understand the reasons for model prediction.

Recently, Large Language Models (LLMs) have shown considerable potential cross various tasks

*The corresponding author.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9127–9142
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

(Zhao et al., 2023). LLMs have pre-trained on extensive data, and leveraging the world knowledge of LLMs for fake news detection has excellent potential. On the one hand, it does not require massive labeled data on specific domain for training and with highly generalization capacity. On the other hand, benefit from the powerful text generation capability of LLMs, it can generate human-readable explanations (Ouyang et al., 2022), which effectively enhanced human confidence in the detection results.

Caramancion (2023) directly prompting LLMs for claim [1] verification, but such a straightforward approach fails to engage the model making detailed reasoning process, resulting in limited capabilities. To enhance the reasoning performance, (Pan et al., 2023; Wang and Shu, 2023; Zhang and Gao, 2023) involve LLMs to decompose claims into simpler sub-claims or sub-questions, and then retrieve external evidences based on these decomposed results. Such step-by-step reasoning process further improved the verification accuracy.

Although these methods effectively utilize LLMs for explainable fake news detection, two key challenges remain: **1) The retrieved evidences are often insufficient.** Existing methods depend on single knowledge source for evidence retrieval, such as specific corpus or online APIs, which will result in incomplete evidence and limit the model's ability to verify claims comprehensively, as illustrated in Figure 1. **2) The retrieved evidence contains substantial redundant information.** Previous methods search for external evidence based on decomposed results, however, the process makes some gap between the original and decomposed claims, introducing irrelevant information during retrieval.

To address these challenges, we propose IMRRF. For the first challenge, we aim for the LLM to provide more knowledge and obtain more comprehensive information through knowledge graph retrieval. For the second challenge, IMRRF bypasses claim decomposition, directly retrieves external evidence from a targeted corpus and refines evidence containing redundant information under the guidance of knowledge graph retrieval results. Finally, we combine the externally retrieved evidence with the supplementary evidence generated by the LLM to verify complex claims. Experiments on two real-world datasets show that IMRRF significantly outperforms existing methods. Our contributions are as follows:

- **Comprehensive Analysis:** We performed an in-depth analysis of existing fact-checking methods that integrate LLMs, identifying the possible reasons causes of their limitations.
- **Innovative Framework:** We propose a novel framework that combines knowledge graph retrieval with corpus-based retrieval, allowing for more comprehensive evidence acquisition. Additionally, we leverage the extensive internal knowledge of LLMs as supplementary evidence for claim verification.
- **Performance Validation:** We validate the efficiency of our approach, demonstrating that it outperforms existing methods in detection accuracy.

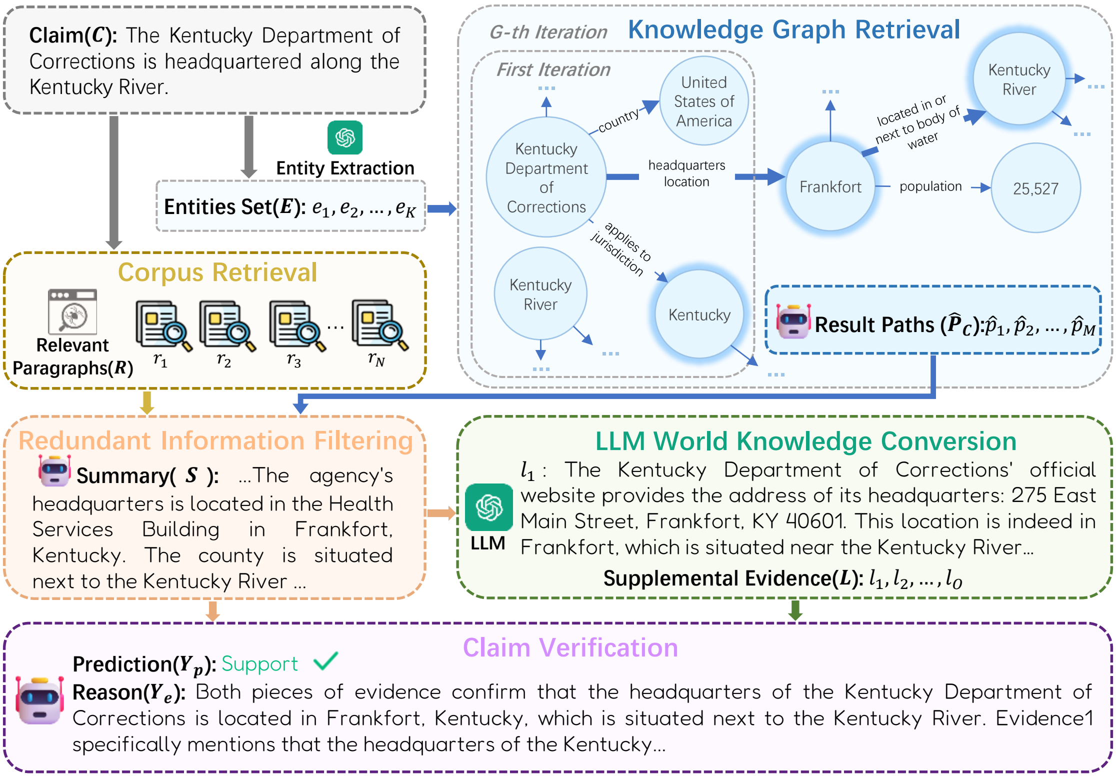
2 Related Work

2.1 Fact-checking

Fact-checking typically involves extracting key information from news content (Konstantinovskiy et al., 2021; Fajcik et al., 2023), followed by retrieving relevant evidence using search engines (Augenstein et al., 2019; Chen et al., 2024; Zhang and Gao, 2023; Wang and Shu, 2023), knowledge graphs(Zou et al., 2023; Liu et al., 2024), and similarity algorithms(Yao et al., 2023). The news content and retrieved evidence are then input into a detection model for verification. Early fact-checking methods required extensive labeled data for model training, and their interpretability was often insufficient during this process.

Recently, LLMs have been applied to this domain. Zhang and Gao (2023) utilize LLMs with few-shot learning to decompose claims into simpler sub-claims, verifying them using the LLMs' internal knowledge when the model is confident; otherwise, they integrate evidence from online sources to complete the verification. Wang and Shu (2023) employ LLMs to decompose claims into first-order logic clauses consisting of predicates, enabling the decomposition of more effective sub-claims, followed by retrieving supporting evidence from the Web and independently verifying each subclaim. Pan et al. (2023) introduce an approach that enables LLMs to adaptively decompose complex claims into multi-step reasoning processes, where the model progressively retrieves external

[1] The news to be verified.


Claim(C): The Kentucky Department of Corrections is headquartered along the Kentucky River.
Entity Extraction
Entities Set(E): $e_1, e_2, \ldots, e_K$
G-th Iteration
Knowledge Graph Retrieval
First Iteration
Kentucky Department of Corrections
country
United States of America
headquarters location
applies to jurisdiction
Kentucky River
Kentucky
Frankfort
located in or next to body of water
Kentucky River
population
25,527
Result Paths ($\hat{P}_C$):$\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_M$
Corpus Retrieval
Relevant Paragraphs(R)
r_1
r_2
r_3
r_N
Redundant Information Filtering
LLM World Knowledge Conversion
LLM
Supplemental Evidence(L): $l_1, l_2, \ldots, l_O$
Claim Verification
Prediction(Y_p): Support ✓

Summary(S): ...The agency's headquarters is located in the Health Services Building in Frankfort, Kentucky. The county is situated next to the Kentucky River ...

l_1 : The Kentucky Department of Corrections' official website provides the address of its headquarters: 275 East Main Street, Frankfort, KY 40601. This location is indeed in Frankfort, which is situated near the Kentucky River...

Reason(Y_e): Both pieces of evidence confirm that the headquarters of the Kentucky Department of Corrections is located in Frankfort, Kentucky, which is situated next to the Kentucky River. Evidence1 specifically mentions that the headquarters of the Kentucky...

Figure 2: Overview of our IMRRF framework, which integrates relevant evidence retrieved from specific corpus and knowledge graph with the extensive internal knowledge of LLM to generate predictive label and rational explanation.

evidence to support inference, ultimately enhancing fact-checking performance. Some researchers bypass the step of claim decomposition and instead directly use claims to search for relevant evidence online (Li et al., 2024a; Chen et al., 2024). Despite these methods exhibiting excellent performance in detection, their reliance on a single knowledge source during the evidence retrieval stage results in insufficient evidence, limiting the model's ability to verify the claims from multiple perspectives comprehensively.

2.2 Knowledge Graph Retrieval

Knowledge graph represents information using nodes, edges, and logical rules, providing structured and explicit knowledge to model (Kim et al., 2023). Currently, few fact-checking methods integrate LLMs with knowledge graph retrieval, as most applications primarily focus on their use for question-answering tasks instead. Sun et al. (2024) enhance the deep reasoning capabilities of LLMs in knowledge-intensive tasks by establishing a foundation for reasoning through the extraction of diverse multi-hop reasoning paths from knowledge graph. Li et al. (2024b) employ CoT reasoning to generate initial reasoning and answers for knowledge-intensive problems. They propose an adaptive query generator to formulate query statements for various types of databases and knowledge bases. Based on these query results, the reasoning steps are gradually refined to derive the final answer. These methods achieve impressive results in question-answering tasks. Therefore, in this work, we explore the application of knowledge graph retrieval in fact-checking approaches that integrate LLM.

3 Methodology

In this paper, we aim to better integrate LLM with fake news detection tasks, producing both final predictions and human-readable explanations. Given a claim C, the summary of external retrieval evidence retrieved from multiple knowledge sources S and the supplementary evidence L generated by LLM, we expect IMRRF to produce a label $Y_p \subseteq \{Supports, Refutes\}$ and human-readable explanation Y_e. Formally, this is expressed as:

$$Y_p, Y_e = IMRRF(C, S, L) \quad (1)$$

As illustrated in Fig 2, we propose a framework that retrieves comprehensive evidence based on claims from multiple knowledge sources. Furthermore, we summarize the evidence to eliminate irrelevant information and use this refined evidence to guide the LLM in converting its extensive internal world knowledge into supplementary evidence. Ultimately, we combine the summarized evidence with the supplementary evidence to generate prediction labels and human-readable explanations. The whole process consists of four stages: multiple knowledge sources retrieval, redundant information filtering, LLM world knowledge conversion, and claim verification.

3.1 Multiple Knowledge Sources Retrieval

Specific Corpus Retrieval: Given a complex claim C, the model retrieves a set of relevant evidence $R = \{r_1, r_2, \ldots, r_N\}$ from a large textual corpus, where N represents the number of retrieved evidence paragraphs. Specifically, for each claim, we employ the BM25 retrieval algorithm (Robertson et al., 1994) to retrieve paragraphs from the Wikipedia corpus.

Key Entity Based Knowledge Graph Retrieval: The model leverages LLM to extract a set of key entities $E = \{e_1, e_2, \ldots, e_K\}$ from C, where e_k represents the k-th entity. In each retrieval iteration g, for every entity e_i^g the model retrieves a set of tuples $K_i^g = \{(e_i^g, r_{i,1}^g, t_{i,1}^g), \ldots, (e_i^g, r_{i,D_i^g}^g, t_{i,D_i^g}^g)\}$ from the knowledge graph. Each tuple follows an entity-relation-target structure, where $t_{i,d}^g$ denotes the target entity connected to e_i^g through the relationship $r_{i,d}^g$, and $d = 1, 2, \ldots, D_i^g$. Here, D_i^g represents the total number of retrieved tuples for e_i^g in iteration g. To refine the retrieved evidence, the model filters the Q most relevant tuples, forming the subset $\hat{K}_i^g$ based on their relevance to C. If the current iteration g has not yet reached the predefined maximum G, the target entities $t_{i,q}^g$ from the selected Q tuples are aggregated to form the entity set for the next retrieval iteration. This iterative process continues until all retrieval steps are completed.

After retrieval concludes, the results are consolidated to construct the knowledge graph path set P_k for each entity e_k. Each path is represented as $p_{k,h} = (e_k, r_{k,h,1}, t_{k,h,1}, \ldots, r_{k,h,Q}, t_{k,h,Q})$, where $t_{k,h,q}$ is the target entity at the q-th hop associated with e_k, and $r_{k,h,q}$ denotes the relationship between $t_{k,h,q-1}$ and $t_{k,h,q}$. Here $h = 1, 2, \ldots, H_k$, and H_k represents the number of knowledge graph paths generated for e_k. To optimize efficiency, the model assesses the retrieved information before querying each entity and at the start of each iteration. If the retrieved knowledge is deemed sufficient to verify the claim, further retrieval is terminated early.

Finally, the knowledge graph path sets of all entities are merged into P_C, and LLM leveraged to select the M most relevant paths to form the optimized knowledge graph path set $\hat{P}_C = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_M)$, completing the extraction and refinement of knowledge graph paths for verification of C. Specifically, we leverage GPT-3.5-Turbo to process the data during retrieval, setting $G = 2$ and $Q \leq 3$, while retrieving information from Wikidata.

3.2 Redundant Information Filtering

After completing the external evidence retrieval process, the model integrates the textual evidence $R = \{r_1, r_2, \ldots, r_N\}$ retrieved from the specific corpus and the set of knowledge graph paths $\hat{P}_C = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_M)$ to generate a refined summary by prompting LLM base on their relevance to C. This process is formally expressed as:

$$S = RIF(C, R, \hat{P}_C) \tag{2}$$

where S denotes the generated summarized evidence, which is obtained by filtering out irrelevant information and refining the retrieved external evidence.

3.3 LLM World Knowledge Conversion

The model takes the summarized evidence S and the claim C as input, prompting the LLM to utilize its extensive internal world knowledge to generate supplementary evidence relevant to C. This process is formally expressed as:

$$L = LWKC(C, S) \tag{3}$$

where $L = \{l_1, l_2, \ldots, l_O\}$ represents the set of supplementary evidence, and O denotes the number of generated evidence items. In our experiments, we set $O = 3$.

3.4 Claim Verification

To determine the final prediction for C, the model integrates the summarized evidence S derived from multiple knowledge sources along with the supplementary evidence $L = \{l_1, l_2, \ldots, l_O\}$ generated by the LLM. The model then performs reasoning over this combined information to predict the label Y_p of the claim and generates a human-readable explanation Y_e.

4 Experiments

4.1 Setting

Dataset. To assess the effectiveness of IMRRF in detecting complex claims, we evaluate its performance on the FEVEROUS(Aly et al., 2021) and HOVER (Jiang et al., 2020) datasets. Compared to common fact-checking datasets, these datasets feature more intricate claims that requires multi-step evidence retrieval and reasoning for validation. The FEVEROUS dataset is designed to validate complex claims involving both structured and unstructured data, with annotations based on sentence and table evidence from Wikipedia. In our experiments, following ProgramFC(Pan et al., 2023), we focus exclusively on claims requiring sentence-based evidence, which results in a subset of 2,962 claims. The HOVER dataset emphasizes multi-hop fact verification, covering claims that necessitate information integration and logical inference across multiple Wikipedia entries. For our experiments, we utilize its validation set, which consists of 1,126 2-hop claims, 1,835 3-hop claims, and 1,039 4-hop claims.

Baseline. We compare the proposed method with the following baselines. ProgramFC(Pan et al., 2023) introduces an approach that enables LLMs to adaptively decompose complex claims into multi-step reasoning processes, progressively retrieving external evidence to support inference. HiSS(Zhang and Gao, 2023) employs LLMs to decompose claims into simpler sub-claims, verifying them using the LLMs' internal knowledge when the model is confident in its verification; otherwise, it integrates evidence from online sources to verify the sub-claims. FOLK(Wang and Shu, 2023) employs LLMs to decompose claims into first-order logic clauses consisting of predicates, enabling the decomposition of more effective sub-claims, retrieving supporting evidence from the Web, and verifying each sub-claim independently. CoK(Li et al., 2024b) utilizes the CoT strategy to generate initial inferences for knowledge-intensive problems, employing an adaptive query generator to construct structured queries for different databases and knowledge bases, and iteratively refining its reasoning steps based on the retrieved results to derive the final answer.

Implementation details. During the external evidence retrieval stage, we utilize GPT-3.5-Turbo[2] to process intermediate steps efficiently. Given its larger parameter scale and enhanced reasoning capabilities, GPT-4o[2] is employed in redundant information filtering and LLM world knowledge conversion steps. For the claim verification phase, we employ FLAN-T5-XL[3] and GPT-4o. Throughout all experiments, the temperature parameter for all LLMs is set to 0 and the full prompt used can be found in the Appendix A.

4.2 Main Results

We report the overall results for IMRRF and the baseline methods in Table 1. Across both datasets, IMRRF achieves higher accuracy, F1 score, and performance in the "Refutes" and "Supports" categories compared to baseline methods, demonstrating its effectiveness. We observe the following:**1)** Compared to HiSS, which employs few-shot prompts to directly decompose the complex claim into sub-claims, FOLK improves sub-claims decomposition by extracting predicates from the claim. ProgramFC further enhances LLMs' reasoning by converting the claim into structured logical reasoning programs. Both FOLK and ProgramFC outperform HiSS in detection performance, highlighting that the quality of claim decomposition directly influences verification performance. However, IMRRF bypasses the claim decomposition step and instead retrieves external evidence directly based on the claim, avoiding limitations with the quality of decomposition while still achieving excellent detection performance. **2)** Compared to HiSS and FOLK, which solely rely on Web retrieval, CoK integrates both Web and knowledge graph retrieval, resulting in superior performance. This demonstrates the advantage of multiple knowledge source retrieval and shows that the knowledge graph can provide more structured and reliable evidence. **3)** On the FEVEROUS dataset, IMRRF outperforms ProgramFC, achieving an 8.38% increase in accuracy and a 9.67% improvement in F1 score when both verification models are FLAN-T5-XL. Similarly, on the HOVER dataset, IMRRF exhibits stronger performance, particularly for complex 3-hop and 4-hop claims, with an average accuracy improvement of 5.9% and an F1 score gain of 7.5% over ProgramFC when both

[2] https://openai.com/

[3] https://huggingface.co/google/flan-t5-xl

Dataset	Method	Accuracy	F1	Refutes			Supports		
				Precision	Recall	F1	Precision	Recall	F1
FEVEROUS	ProgramFC	66.64	65.21	63.98	83.04	72.28	72.29	48.62	58.14
	HiSS	61.11	60.24	70.50	44.23	54.36	56.51	**79.66**	66.12
	FOLK	62.82	61.42	61.34	78.28	68.78	65.81	45.86	54.05
	CoK	63.77	63.64	69.46	55.00	61.39	59.75	73.42	65.88
	IMRRF(FLAN-T5-XL)	**75.02**	**74.88**	**74.92**	78.59	76.71	75.13	71.08	**73.05**
	IMRRF(GPT-4o)	72.42	71.07	67.90	**89.75**	**77.31**	**82.57**	53.37	64.83
HOVER 2-hop	ProgramFC	72.29	72.06	73.52	75.70	74.59	70.78	68.33	69.53
	HiSS	59.86	59.85	64.35	56.69	60.28	55.82	63.53	59.43
	FOLK	61.73	58.86	60.61	82.06	69.72	64.71	38.15	48.00
	CoK	60.83	60.46	63.02	65.62	64.29	58.06	55.25	56.64
	IMRRF(FLAN-T5-XL)	73.89	73.89	**80.31**	68.10	73.70	68.52	**80.61**	**74.07**
	IMRRF(GPT-4o)	**75.67**	**75.54**	77.45	**77.19**	**77.32**	**73.61**	73.90	73.75
HOVER 3-hop	ProgramFC	60.71	60.19	56.20	76.36	64.74	68.80	46.69	55.63
	HiSS	55.59	55.29	52.33	67.47	58.94	60.67	44.94	51.63
	FOLK	53.69	50.34	50.70	**83.99**	63.23	64.62	26.36	37.44
	CoK	57.10	56.53	53.49	72.43	61.54	63.57	43.30	51.51
	IMRRF(FLAN-T5-XL)	66.92	66.80	**65.15**	64.48	64.81	68.47	**69.11**	**68.79**
	IMRRF(GPT-4o)	**67.30**	**67.11**	62.04	79.35	**69.64**	**75.34**	56.51	64.58
HOVER 4-hop	ProgramFC	57.36	54.92	55.64	79.36	65.42	61.89	34.64	44.42
	HiSS	53.71	52.52	53.48	68.37	60.02	54.12	38.55	45.03
	FOLK	54.81	48.35	53.09	**89.40**	66.62	64.52	19.61	30.08
	CoK	56.76	55.32	55.68	73.43	63.34	58.94	39.49	47.29
	IMRRF(FLAN-T5-XL)	63.04	**63.01**	**63.33**	64.77	64.04	62.73	**61.25**	**61.98**
	IMRRF(GPT-4o)	**63.33**	61.74	60.17	82.29	**69.54**	**70.57**	43.64	53.93

Table 1: The accuracy, F1 score, and detailed precision, recall, and F1 score for the "Refutes" and "Supports" categories of IMRRF and the baseline models on the evaluation sets of the FEVEROUS and HOVER datasets.

verification models are FLAN-T5-XL. These results highlight that IMRRF effectively integrates specific textual corpus and knowledge graph evidence while leveraging the internal knowledge of LLMs for supplementary evidence, leading to more comprehensive and precise claim verification, enabling the model to achieve superior performance. **4)** On the HOVER 2-hop dataset, methods such as HiSS, FOLK, and CoK exhibit similar performance, with even the best baseline, ProgramFC, showing minimal improvement over IMRRF. This occurs because, for simpler claims, the retrieved evidence across different methods remains largely consistent, resulting in small variations in detection performance. However, when handling more complex claims, IMRRF provides more comprehensive and accurate evidence, leading to a clear advantage in detection performance. **5)** On the FEVEROUS dataset, IMRRF demonstrates a well-balanced precision, recall, and F1 score across both the "Refutes" and "Supports" categories, achieving a significantly higher F1 score than the compared methods. This indicates that IMRRF effectively retrieves diverse evidence while maintaining high accuracy. Similarly, on the HOVER dataset, IMRRF maintains an advantage in the "Refutes" category and continues to excel in handling 3-hop and 4-hop claims in the "Supports" category, showing stable recall performance and strong adaptability in verifying complex claims. **6)** On the FEVEROUS dataset, when IMRRF employs GPT-4o as the verification model, its detection performance may be slightly lower than that of FLAN-T5-XL due to hallucination issues inherent in large language models. However, on the HOVER dataset, GPT-4o exhibits stronger reasoning capabilities, particularly in handling complex multi-hop claims, allowing it to outperform FLAN-T5-XL.

4.3 Ablation Study

To analyze the contribution of each component in IMRRF, we utilize FLAN-T5-XL as the verification model and conduct ablation experiments. The results are shown in Figure 3. First, we evaluate the impact of incorporating knowledge graph retrieval alongside textual corpus retrieval. This enhancement improves accuracy by 0.4% on the FEVEROUS dataset and by 1.34% on the HOVER 2-hop dataset. However, on the 3-hop and 4-hop

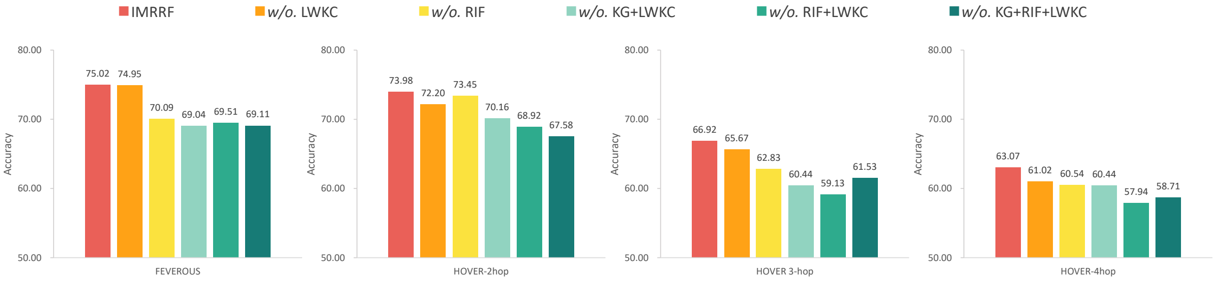


Figure 3: Ablation results of IMRRF on the FEVEROUS and HOVER datasets, where KG denotes knowledge graph retrieval, RIF denotes redundant information filtering, and LWKC denotes LLM world knowledge conversion.

datasets, accuracy decreases by 2.4% and 0.77%, respectively. These results suggest that while integrating multiple knowledge sources of evidence can strengthen model verification, the model also struggles to filter out irrelevant information, which negatively impacts verification performance.

Next, we analyze the role of redundant information filtering. Compared to summarizing evidence solely from the textual corpus, integrating knowledge graph retrieval before summarization leads to accuracy improvements of 1.05%, 3.29%, 0.39%, and 0.1% across the datasets. These results highlight the effectiveness of knowledge graph enhance retrieval by providing additional relevant evidence, while demonstrating that redundant information filtering plays a crucial role in refining retrieved evidence by eliminating irrelevant information and improving the overall quality of evidence.

Furthermore, we leverage the retrieved textual corpus and knowledge graph evidence to guide LLM directly in converting its internal knowledge into supplementary evidence and evaluate its impact as an additional input. On the FEVEROUS dataset, incorporating LLM-generated supplementary evidence improves accuracy by 3%. Similarly, on the HOVER dataset, accuracy increases by 3.28%, 6.54%, and 3.08% for HOVER 2-hop,3-hop, and 4-hop, respectively. These results demonstrate the capability of LLM to contribute valuable supplementary evidence, especially in more complex multi-hop verification tasks.

Finally, we leverage summarized evidence from both textual corpus and knowledge graph retrievals enhanced by redundant information filtering to further guide the LLM in world knowledge conversion. This approach yields an accuracy improvement of 4.93% on FEVEROUS and provides additional gains of 0.53%, 4.09%, and 2.53% across different levels of claim complexity in the HOVER dataset. These results reaffirm the importance of redundant information filtering and LLM world knowledge conversion in claim verification.

4.4 Cross Model Evaluation

In addition to conducting experiments on redundant information filtering and LLM world knowledge conversion using GPT-4o, we also evaluate the performance of IMRRF with other LLMs, including GPT-3.5-Turbo, Meta Llama3-70B[4], and Gemini-1.0-Pro[5], to validate the effectiveness of IMRRF. In all experiments, FLAN-T5-XL is used as the verification model.

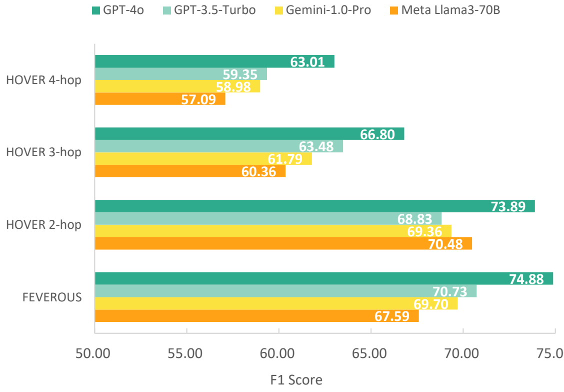


Figure 4: The performance of IMRRF across different LLMs in redundant information filtering and world knowledge conversion.

As shown in Figure 4, IMRRF consistently demonstrates strong detection performance across all LLMs, even achieving results on par with the best baseline method, ProgramFC, when using the relatively weaker Meta Llama3-70B. These results confirm the robustness and effectiveness of IMRRF across diverse LLMs. Notably, the performance of LLMs, aside from GPT-4o, is slightly inferior to ProgramFC on the HOVER 2-hop dataset. As

[4] https://huggingface.co/meta-llama/Meta-Llama-3-70B

[5] https://gemini.google.com

discussed previously, for simpler claims like 2-hop, the evidence retrieved by various methods tends to be similar, resulting in minimal differences in detection performance across models.

4.5 LLM Evidence Quality Evaluation

To evaluate the quality of internal knowledge conversion across different LLMs when handling complex claims, we employ the same LLM for redundant information filtering and use the summarized evidence to guide different LLMs in world knowledge conversion. Specifically, we employ GPT-3.5-Turbo for the redundant information filtering task. In all experiments, FLAN-T5-XL is used as the verification model.

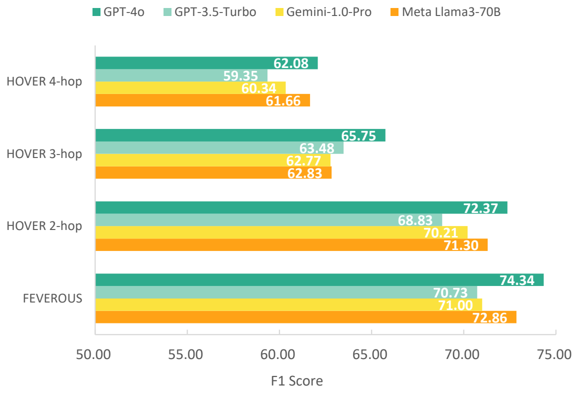


Figure 5: The performance of using the same summarized evidence to guide different LLMs in world knowledge conversion and generate supplementary evidence as additional input.

As shown in Figure 5, when using the same summarized evidence, Meta Llama3-70B, despite having fewer parameters than GPT-3.5-Turbo and Gemini-1.0-Pro, outperforms both GPT-3.5-Turbo and Gemini-1.0-Pro in generating supplementary evidence for complex claims. These results illustrate the importance of the quality of the summary after redundant information filtering for LLM world knowledge conversion. As expected, GPT-4o, with its extensive parameters and precise internal knowledge, achieves the best detection performance.

4.6 Prompt Sensitivity Evaluation

To assess the impact of different prompts on performance in the LLM world knowledge conversion, we rewrite the original prompt in two ways: Type I modifies the task description, and Type II abbreviates the original prompt. As shown in Table 2, as long as the prompt is clearly stated, variations in phrasing do not significantly affect the quality of evidence generation by the LLM.

Dataset	Prompt Type		
	Original	Type-I	Type-II
FEVEROUS	**75.02**	74.24	73.18
HOVER 2-hop	73.98	**75.22**	74.60
HOVER 3-hop	66.92	**67.14**	65.17
HOVER 4-hop	63.07	61.60	**63.11**

Table 2: F1 score of different prompts in the LLM world knowledge conversion on GPT-4o.

4.7 Error Analysis

We manually analyze 50 error cases randomly selected from the FEVEROUS and HOVER datasets, respectively, using GPT-4o as the validation model. We classify the errors into three categories: **1)** *Insufficiency of evidence*, the multiple knowledge sources retrieval and LLM world knowledge conversion fail to provide sufficient evidence to comprehensively validate the claim; **2)** *Hallucinated evidence*, the LLM world knowledge conversion produces hallucinatory evidence, directly impacting claim validation; **3)** *Verification hallucination*, although the retrieved and converted knowledge provides sufficient evidence, the model deviates or hallucinates during the validation step, leading to incorrect conclusions.

Error Type	Proportion(%)	
	FEVEROUS	HOVER
Insufficiency of evidence	**50**	32
Hallucinated evidence	22	**52**
Verification of hallucination	28	16

Table 3: The proportion of different types of error cases in the FEVEROUS and HOVER datasets.

As shown in Table 3, previous experiments have demonstrated that IMRRF already performs well on the FEVEROUS dataset. However, most errors stem from insufficiency of evidence, which is the primary cause of claim validation failure. On the HOVER dataset, which requires multi-hop reasoning and contains relatively complex retrieved evidence, the LLM world knowledge conversion is more prone to hallucinations. Thus, on the HOVER dataset, errors are primarily caused by hallucinated evidence. Each type of error case can be found in Appendix C.

5 Conclusion

This paper proposes IMRRF, which retrieves comprehensive evidence based on claims from multiple knowledge sources. In the retrieval process, IMRRF integrates specific textual corpus retrieval with knowledge graph retrieval to obtain more comprehensive evidence. Subsequently, the model summarizes the retrieved external evidence and filters out irrelevant information to ensure the reliability of the evidence. Using the summarized evidence, the model guides the LLM to leverage its extensive internal world knowledge to generate supplementary evidence, thereby verifying the claims more effectively. IMRRF demonstrates promising performance on the HOVER and FEVEROUS datasets, and we validate its effectiveness across different LLMs, with results demonstrating its superiority. Additionally, we examine the quality of supplementary evidence generated by different LLMs. We also conduct experiments on prompt sensitivity and analyze error cases. Our research introduces a novel detection method for fact-checking and contributes to the development of more effective verification approaches.

Limitations

Despite IMRRF demonstrating exceptional performance in detecting complex claims, there are areas for improvement:**1)** In the knowledge graph retrieval phase, extracting all available entities from the claim may lead to excessive retrieval attempts. **2)** Currently, our external evidence retrieval primarily relies on specific corpus and knowledge graph, while Web search remains an underexplored option. Future work could investigate incorporating Web search to further optimize the model.

Ethical Considerations

Biases. We note that there might be some biases in the data used to train the LLMs, as well as in factuality judgments. Both are beyond our control.

Intended Use and Misuse Potential. Our models can be of interest to the general public and could also save a lot of time to human fact-checkers.

Environmental Impact. The use of large language models requires a significant amount of energy for computation for training, which contributes to global warming.

Acknowledgements

This work is supported by the Yunnan Province expert workstations (Grant No: 202305AF150078), National Natural Science Foundation of China (Grant No: 62162067, 62462067), Yunnan Fundamental Research Project (Grant No: 202401AT070474), Yunnan Province Special Project (Grant No:202403AP140021), and Scientific Research and Innovation Project of Postgraduate Students in the Academic Degree of Yunnan University.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification, FEVER@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 85–90. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4684–4696. Association for Computational Linguistics.

Kevin Matthe Caramancion. 2023. News verifiers showdown: A comparative performance evaluation of chatgpt 3.5, chatgpt 4.0, bing ai, and bard in news fact-checking. *CoRR*, abs/2306.17176.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3569–3587. Association for Computational Linguistics.

Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. 2020. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *WWW '20:*

The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020, pages 2892–2898. ACM / IW3C2.

Limeng Cui, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: A system for explainable fake news detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 2961–2964. ACM.

Martin Fajcik, Petr Motlícek, and Pavel Smrz. 2023. Claim-dissector: An interpretable fact-checking system with joint re-ranking and veracity prediction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10184–10205. Association for Computational Linguistics.

Joonwon Jang, Yoon-Sik Cho, Minju Kim, and Misuk Kim. 2022. Detecting incongruent news headlines with auxiliary textual information. *Expert Syst. Appl.*, 199:116866.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Kumar Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3441–3460. Association for Computational Linguistics.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16190–16206. Association for Computational Linguistics.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *Digital Threats*, 2(2).

Guanghua Li, Wensheng Lu, Wei Zhang, Defu Lian, Kezhong Lu, Rui Mao, Kai Shu, and Hao Liao. 2024a. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. *CoRR*, abs/2403.09747.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024b. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024. TELLER: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15556–15583. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3818–3824. IJCAI/AAAI Press.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6981–7004. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3346–3359. Association for Computational Linguistics.

Tina Esther Trueman, Ashok Kumar J., Narayanasamy P., and Vidya J. 2021. Attention-based c-bilstm for fake news detection. *Appl. Soft Comput.*, 110:107600.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6288–6304. Association for Computational Linguistics.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2733–2743. ACM.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3901–3907. ijcai.org.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, IJCNLP 2023 -Volume 1: Long Papers, Nusa Dua, Bali, November 1 - 4, 2023*, pages 996–1011. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

Anni Zou, Zhuosheng Zhang, and Hai Zhao. 2023. Decker: Double check with heterogeneous knowledge for commonsense fact verification. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11891–11904. Association for Computational Linguistics.

A Prompt

The prompts used in the IMRRF reported in the following tables respectively. The {·} in prompt represent the input content.

```
[Guidance]
Given a claim, if I want to verify the truth
or falseness of the claim, help me extract the
entities of the claim to be more suitable for
knowledge graph to search for evidence.  The
entities should be short and no more than 5.
Only entities such as proper names, places, and
person need to be extracted, ignoring entities
such as time, data, numbers, country, and verbs.
If there is no suitable entity just answer None.
[Input]
Claim: {claim}
[Output Format]
Entities: [Entity 1, Entity 2, ...] Remember to
follow the format for output.
```

Table 4: Prompt of *Extracting Key Entities From Claim* in our experiments.

```
[Guidance]
Based on the results retrieved from the following
knowledge graph, choose some of the most relevant
path to claim and only return their indices:
[Input]
Claim: {claim}
Results: {result chain}
[Output Format]
Idx: [idx1, idx2, ...] Remember to follow the
format for output.
```

Table 5: Prompt of *Related Result Path Selection* in our experiments.

```
[Guidance]
As a paragraph-summarizing assistant, you are
required to complete the task according to the
following rules: 1. Extract information related
to the claim from the provided paragraphs and
summarize it; the summary should not exceed 300
words. 2. Summarize all the information in the
paragraphs that is relevant to the claim, but
do not generate a summary based directly on the
claim, as the claim may be incorrect. 3. Summary
should be accurate and comprehensive.  4.  Do
not summarize irrelevant information.  5.  Do
not generate information that is not relevant to
summary.
[Input]
Claim: {claim}
Paragraph1: {paragraph1}
Paragraph2: {paragraph2}
```

Table 6: Prompt of *Redundant Information Filtering* in our experiments.

```
[Guidance]
As an Information Retrieval Assistant, you
are required to complete the task according
to the following rules:  1.  Based on the
given information, retrieve 3 additional pieces
of information that can help determine the
correctness of the claim.  2.  Each piece of
additional information should not exceed 100
words. 3. Do not generate additional information
directly based on the claim, as the claim may
be incorrect.  4.  If there is an error in the
given information, please provide the correct
information as additional information.
[Input]
Claim: {claim}
Summary: {summary}
[Output Format]
Additional information:
```

Table 7: Prompt of *LLM World Knowledge Conversion* in our experiments.

```
[Guidance]
You are a CLAIM VERIFICATION ASSISTANT and you
need to determine if the claim is correct based
on the given evidence.
[Input]
Claim: {claim}
Evidence1: {summary}
Evidence2: {llm evidence}
Based on the above evidences, is it true that
claim?  Please provide the answer (true/false)
and the reason.
[Output Format]
answer:
reason:
```

Table 8: Prompt of *Claim Verification* in our experiments.

```
[Guidance]
As a knowledgeable robot assistant, use the
provided clues to generate three concise pieces
of evidence (each under 100 words) to help verify
a claim. Be cautious, as the initial information
may be inaccurate. Ensure each piece of evidence
is based on factual data.
[Input]
Claim: {claim}
Summary: {summary}
[Output Format]
Additional information:
```

Table 9: Rewriting the prompt for Type-I in *LLM World Knowledge Conversion*.

```
[Guidance]
As an Info Retrieval Assistant: 1.Retrieve 3
extra pieces from your knowledge to help verify
claim’s correctness. 2.Each piece under 100
words. 3.Don’t generate info from claim (it may
be wrong). 4.If given info has error, provide
correct info as extra.
[Input]
Claim: {claim}
Summary: {summary}
[Output Format]
Additional information:
```

Table 10: Rewriting the prompt for Type-II in *LLM World Knowledge Conversion*.

B Case Study

In Table 11 and Table 12, we report two cases from HOVER and FEVEROUS datasets. As illustrated, by retrieving information from multiple knowledge sources and leveraging the world knowledge of LLM, IMRRF is capable of gathering more comprehensive evidence. Additionally, the strategy of redundant information filtering helps eliminate many irrelevant details, which enables the model to make accurate predictions about the labels of complex claims and provide reasonable explanations.

C Error Analysis

In Tables 13 - 15, we report the real examples for each error type.

Example 1
Claim: The Kentucky Department of Corrections is headquartered along the Kentucky River.
Specific Corpus Retrieval: 1. The Kentucky Department of Juvenile Justice is a state agency of Kentucky headquartered in unincorporated Franklin... 2. Kentucky Correctional Institution for Women is a prison located in unincorporated Shelby County... 3. The Kentucky State Penitentiary, also known as the "castle on the Cumberland," is a maximum security ... 4. The Kentucky Department of Corrections is a state agency of the Kentucky Justices... The agency is headquartered in the Health Services Building in Frankfort... 5. Little Sandy Correctional Complex is a minimum and medium/maximum-security prison located in Elliott County...
Key Entity Based Knowledge Graph Retrieval: 1. The Kentucky Department of Corrections has its headquarters located in Frankfort, which is a county seat city in Franklin County, Kentucky, and is situated next to the Kentucky River. 2. The Kentucky River is located in Kentucky, and the state of Kentucky is named after the Kentucky River.
Redundant Information Filtering: 1. The Kentucky Department of Corrections is a state agency under the Kentucky Justice & Public Safety Cabinet... 2. The agency's headquarters is located in the Health Services Building in Frankfort, Kentucky. Frankfort, the county seat of Franklin County, is situated next to the Kentucky River.
LLM World Knowledge Conversion: 1. The Kentucky Department of Corrections' official website provides the address of its headquarters: 275 East Main Street, Frankfort, KY 40601. This location is indeed in Frankfort, which is situated near the Kentucky River. 2. According to the Franklin County Property Valuation Administrator's office, the area where the Kentucky Department... 3. The Kentucky River runs through Frankfort, the capital city of Kentucky, and many state government buildings..
Claim Verification: Prediction: Support Reason: Both pieces of evidence confirm that the headquarters of the Kentucky Department of Corrections is located in Frankfort, Kentucky, which is situated next to the Kentucky River. Therefore, it is true that the Kentucky Department of Corrections is headquartered along the Kentucky River.

Table 11: Example of an intermediate process output for IMRRF verification of a complex claim.

Example 2
Claim: Ron DeSantis never ran in the 2018 Florida gubernatorial election.
Specific Corpus Retrieval: 1. 2018 Florida gubernatorial election The 2018 Florida gubernatorial election was held on November 6, 2018... 2. 2022 Florida gubernatorial election The 2022 Florida gubernatorial election will take place on November 8, 2022... 3. Ron DeSantis Ronald Dion DeSantis (born September 14, 1978) is an American attorney and politician serving as the 46th and current List of governors of Florida|governor of Florida... 4. Adam Putnam Adam Hughes Putnam (born July 31, 1974) is an American politician and a... 5. Marcellus Osceola Jr.Marcellus William Osceola Jr. (born in 1972) is the current and 7th Tribal Council Chairman...
Key Entity Based Knowledge Graph Retrieval: 1.Ron DeSantis, also known as Ronald Dion DeSantis, has been the Governor of Florida since 2019 and was a successful candidate in the 2018 Florida Governor election. 2. Ron DeSantis was a candidate in the 2018 Florida Governor election, which was a race between him and...
Redundant Information Filtering: Ron DeSantis ran in the 2018 Florida gubernatorial election. The election was held on November 6, 2018, to elect the next Governor of Florida. DeSantis won the Republican primary on August 28, 2018, defeating Adam Putnam, who was initially considered the frontrunner...
LLM World Knowledge Conversion: 1. **Election Results**: According to the Florida Department of State, Division of Elections, Ron DeSantis won the 2018 Florida gubernatorial election with 49.6% of the vote, defeating Democratic candidate Andrew Gillum, who received 49.2% of the vote... 2. **Primary Election**: The New York Times reported that Ron DeSantis won the Republican primary for the 2018 Florida gubernatorial election on August 28, 2018, defeating Adam Putnam with 56.5% of the vote... 3. **Inauguration**: Ron DeSantis was inaugurated as the 46th Governor of Florida on January 8, 2019...
Claim Verification: Prediction: Refute Reason: The evidence clearly indicates that Ron DeSantis did run in the 2018 Florida gubernatorial election. He won the Republican primary on August 28, 2018, and subsequently faced Democratic nominee Andrew Gillum in the general election held on November 6, 2018. DeSantis won the general election and was inaugurated as the 46th Governor of Florida on January 8, 2019. Therefore, the claim that Ron DeSantis never ran in the 2018 Florida gubernatorial election is false.

Table 12: Example of an intermediate process output for IMRRF verification of a complex claim.

Insufficiency of evidence
Claim: Clydebuilt is on the south bank of the River Clyde.
Redundant Information Filtering: The Clydebuilt maritime museum is located in Braehead, which is situated in Renfrewshire. Tradeston and Glasgow Central Mosque are both located on the south bank of the River Clyde in Glasgow. The Renfrew Ferry links the north and south banks of the River Clyde, crossing between Renfrew and Yoker.
LLM World Knowledge Conversion: 1. **Braehead Location**: Braehead is a shopping and leisure complex located in Renfrew, Renfrewshire, Scotland. It is situated on the north bank of the River Clyde, opposite the town of Clydebank. This information helps clarify the location of the Clydebuilt maritime museum within Braehead. 2. **Renfrewshire Geography**: Renfrewshire is a historic county and council area in the west central Lowlands of Scotland. It is bordered to the north by the River Clyde. This geographical context can help determine the relative positions of various landmarks and confirm the bank on which they are located. 3. **Clydebuilt Maritime Museum**: The Clydebuilt maritime museum, which was previously located in Braehead, Renfrewshire, has been closed since 2010. The museum was dedicated to the shipbuilding heritage of the River Clyde. This historical context can help verify the current status and past location of the museum.
Verification: Prediction: Refute Reason: According to the provided evidence, the Clydebuilt maritime museum was located in Braehead, which is situated on the north bank of the River Clyde in Renfrew, Renfrewshire, Scotland. Therefore, it is not on the south bank of the River Clyde.

Table 13: Example of a prediction error caused by insufficient evidence.

Hallucinated evidence
Claim: The star of the Irish film directed by Paddy Breathnach played Marcus Agrippa in the HBO drama series Rome.
Redundant Information Filtering: Alan Leech, an Irish actor, is known for his role as Tom Branson in Downton Abbey and played Marcus Agrippa in the HBO historical drama series Rome. He also starred in the 2004 Irish comedy film Man About Dog, which was directed by Paddy Breathnach. Breathnach directed several other films, including Viva, a 2015 Spanish-language Irish drama set in Cuba, Shrooms, a 2007 horror film, and Ailsa, a 1994 film that won an award at the San Sebastián International Film Festival.
LLM World Knowledge Conversion: 1. **Alan Leech's Filmography**: Alan Leech is an Irish actor known for his roles in Downton Abbey and Bohemian Rhapsody. However, he did not play Marcus Agrippa in the HBO series Rome. The role of Marcus Agrippa was played by Allen Leech, who is often confused with Alan Leech due to the similarity in their names. 2. **Paddy Breathnach's Directorial Works**: Paddy Breathnach is an Irish filmmaker known for directing Viva (2015), Man About Dog (2004), Shrooms (2007), and Ailsa (1994). None of these films feature Alan Leech in a role that would connect him to the character Marcus Agrippa. 3. **HBO Series Rome Cast**: The HBO series Rome featured a cast that included Kevin McKidd as Lucius Vorenus, Ray Stevenson as Titus Pullo, and Simon Woods as Gaius Octavian. The role of Marcus Agrippa was played by Allen Leech, not Alan Leech. This distinction is crucial in verifying the accuracy of the claim.
Verification: Prediction: Refute Reason: The claim is false because the role of Marcus Agrippa in the HBO series Rome was played by Allen Leech, not Alan Leech. The evidence clarifies that Alan Leech, who starred in the Irish film Man About Dog directed by Paddy Breathnach, is often confused with Allen Leech due to the similarity in their names. Therefore, the star of the Irish film directed by Paddy Breathnach did not play Marcus Agrippa in the HBO drama series Rome.

Table 14: Example of a prediction error caused by hallucinated evidence.

Verification hallucination
Claim: The american sitcom with 247 episodes that Jeff Mooring appeared in was Jeff Mooring.
Redundant Information Filtering: Jeff Mooring is an American actor best known for his role as Dave on the television series Sports Night. He has also appeared on other TV shows such as The West Wing, The Nanny, Murphy Brown, and The Cosby Show. There is no mention of him appearing in a sitcom with 247 episodes named Jeff Mooring.
LLM World Knowledge Conversion: 1. **Jeff Mooring's Filmography**: To verify the claim, it is essential to review Jeff Mooring's complete filmography. This can be found on databases like IMDb, which lists all the television series and movies an actor has appeared in, along with the number of episodes for each series. 2. **Sitcoms with 247 Episodes**: Research which American sitcoms have exactly 247 episodes. This can help cross-reference whether Jeff Mooring appeared in any of these shows. Popular databases like TV Guide or Wikipedia can provide lists of sitcoms and their episode counts. 3. **Jeff Winger and Community**: The summary mentions Jeff Winger from Community and an episode involving his father. Confirming the actors involved in Community and their episode counts can help clarify any confusion between Jeff Mooring and Jeff Winger. This information can be found on the official Community series page on NBC or streaming platforms like Netflix.
Verification: Prediction: Refute Reason: The evidence provided does not support the claim that Jeff Mooring appeared in an American sitcom with 247 episodes named Jeff Mooring. Jeff Mooring is an actor known for his roles in various TV shows, but there is no mention of him appearing in a sitcom with 247 episodes. Additionally, there is no sitcom named Jeff Mooring. The evidence suggests that there might be confusion between the actor's name and the title of a show, but no such show exists.

Table 15: Example of a prediction error caused by verification hallucination.