Resolving Conflicting Evidence in Automated Fact-Checking: A Study on Retrieval-Augmented LLMs

Ziyu Ge[1], **Yuhao Wu**[1], **Daniel Wai Kit Chin**[1], **Roy Ka-Wei Lee**[1] and **Rui Cao**[2]
[1]Singapore University of Technology and Design
[2]University of Cambridge
{ziyu_ge, roy_lee}@sutd.edu.sg, {yuhao_wu, daniel_chin}@mymail.sutd.edu.sg, rc990@cam.ac.uk

Abstract

Large Language Models (LLMs) augmented with retrieval mechanisms have demonstrated significant potential in fact-checking tasks by integrating external knowledge. However, their reliability decreases when confronted with conflicting evidence from sources of varying credibility. This paper presents the first systematic evaluation of Retrieval-Augmented Generation (RAG) models for fact-checking in the presence of conflicting evidence. To support this study, we introduce **CONFACT** (**Con**flicting Evidence for **Fact**-Checking), a novel dataset comprising questions paired with conflicting information from various sources. Extensive experiments reveal critical vulnerabilities in state-of-the-art RAG methods, particularly in resolving conflicts stemming from differences in media source credibility. To address these challenges, we investigate strategies to integrate media background information into both the retrieval and generation stages. Our results show that effectively incorporating source credibility significantly enhances the ability of RAG models to resolve conflicting evidence and improve fact-checking performance.

1 Introduction

Motivation. Fact-checking systems are essential tools for combating the spread of misinformation, as they help verify claims by retrieving and analyzing evidence from diverse sources [Guo *et al.*, 2022; Nakov *et al.*, 2021]. Modern fact-checking pipelines increasingly rely on Retrieval-Augmented Generation (RAG) frameworks, which integrate external evidence into Large Language Models (LLMs) to verify claims [Lewis *et al.*, 2020; Guu *et al.*, 2020]. However, a critical challenge arises when fact-checking systems encounter *conflicting evidence*—that is, when retrieved documents present opposing stances on a claim, often originating from sources with varying levels of credibility [Guo *et al.*, 2022; Schlichtkrull, 2024; Hong *et al.*, 2024].

For example, consider the claim: "*Paul Pogba retired from international football in response to French President Macron's comments on Islamist terrorism*", the retrieved evidence might include conflicting documents, as shown in Figure 1, such as one from BBC[1], a highly credible source, and another from Mehr News Agency[2], which is flagged as untrustworthy[3]. To fact-check this claim accurately, a fact-checking system must not only analyze the evidence but also assess the credibility of each source—prioritizing reliable information while discounting less trustworthy content.

Figure 1: The retrieved documents from Google to verify the claim. The retrieved documents from different media sources have different stances towards the claim.

This challenge is exacerbated by the rapid proliferation of low-credibility content and automated misinformation generated by LLMs themselves [Chen and Shu, 2024; Wang *et al.*, 2024a]. Fact-checking in this context requires robust systems capable of resolving conflicts in evidence while reasoning about source credibility—capabilities that are currently

[1]https://www.bbc.co.uk/sport/football/54691842
[2]https://en.mehrnews.com/news/165168/Pogba-quits-intl-football-after-comments-from-Macron-report
[3]https://mediabiasfactcheck.com/mehr-news-agency/

underexplored in fact-checking research.

Research Objectives. Addressing these gaps, this paper focuses on the problem of *fact-checking with conflicting evidence*, where retrieved documents present opposing stances on a claim. Specifically, we aim to evaluate the ability of retrieval-augmented LLMs to identify, analyze, and resolve conflicts in evidence by determining which sources to trust for claim verification. To enable this, we introduce CONFACT (**Con**flicting Evidence for **Fact**-Checking), a novel dataset designed to systematically study this challenge. Each instance in CONFACT comprises a claim paired with documents exhibiting conflicting stances, annotated with source credibility ratings.

We conduct extensive experiments to evaluate state-of-the-art RAG models on CONFACT, revealing critical limitations in their ability to reason through conflicting evidence and prioritize trustworthy sources. Motivated by these findings, we further explore strategies for incorporating media background information—such as source metadata and credibility scores—into both the retrieval and generation processes. Our results demonstrate that effectively integrating source credibility enhances the robustness of retrieval-augmented LLMs in resolving conflicting evidence for fact-checking.

Contributions. In this work, we made the following contributions in this work:

- **Dataset Creation**: We introduce CONFACT, a novel dataset for studying fact-checking with conflicting evidence.[4] The dataset includes claims paired with conflicting retrieved documents, annotated with source credibility and stance labels to facilitate systematic evaluation.
- **Performance Evaluation**: We conduct a comprehensive evaluation of RAG-based LLMs on CONFACT, revealing critical vulnerabilities in resolving conflicting evidence and reasoning about source credibility.
- **Methodological Innovations**: We propose and evaluate multiple strategies for integrating media background information into RAG pipelines, demonstrating significant improvements in fact-checking performance through effective credibility-aware reasoning.

2 Related Work

2.1 RAG for Automated Fact-Checking

Automated fact-checking (AFC) has gained significant attention in recent years [Guo *et al.*, 2022; Nakov *et al.*, 2021]. While LLMs have demonstrated strong performance in various Natural Language Understanding (NLU) tasks [Li *et al.*, 2023], they remain limited in AFC, as fact-checking often requires evidence beyond the parametric knowledge stored within LLMs [Schlichtkrull *et al.*, 2023; Thorne *et al.*, 2018; Wang, 2017]. RAG [Lewis *et al.*, 2020; Ram *et al.*, 2023] facilitates the adaptation of LLMs to AFC by incorporating external retrieved evidence to LLMs [Pan *et al.*, 2023a; Pan *et al.*, 2023b; Chen *et al.*, 2024; Zhang and Gao, 2024]. However, not all retrieved evidence is reliable [Guo *et al.*, 2022; Hong *et al.*, 2024], and information from untrustworthy sources may contain misinformation, leading to conflicting evidence. Recent studies have shown that retrieval-augmented LLMs are particularly vulnerable to contradictions in augmented texts [Min *et al.*, 2020; Lee *et al.*, 2024; Chen *et al.*, 2021; Amplayo *et al.*, 2023]. Given the risks posed by unreliable sources, it is crucial to investigate the robustness of retrieval-augmented LLMs in AFC, particularly in handling conflicting evidence.

2.2 Source Credibility Estimation

Source credibility estimation is crucial, as not all media sources are reliable; however, this problem remains underexplored. Early works addressed this issue by estimating media credibility through analysis of fake news records associated with sources [Mukherjee and Weikum, 2015; Popat *et al.*, 2016; Popat *et al.*, 2017]. The authors in [Baly *et al.*, 2018] introduced the first dataset with human-annotated factuality ratings of news sources and utilized various features, such as Wikipedia information and source URLs, for credibility estimation. Subsequent studies proposed more robust models using diverse features of media sources [Zhang *et al.*, 2019; Baly *et al.*, 2020; Hounsel *et al.*, 2020]. In contrast to these classification approaches, the work in [Schlichtkrull, 2024] emphasized the generation of detailed background checks for media sources.

Despite these advancements, the impact of estimated source credibility in fact-checking is still unknown. has received limited attention. To date, only [Schlichtkrull, 2024] conducted a small-scale experiment with 20 claims, examining whether incorporating source background checks could benefit claim verification. In this paper, we extend this line of inquiry by comprehensively evaluating how media source backgrounds can facilitate fact-checking models, and exploring optimal strategies for integrating source credibility information into these systems.

3 CONFACT Dataset

The CONFACT dataset is specifically designed to facilitate the study of fact-checking in scenarios where conflicting evidence is retrieved from sources of varying credibility, thereby addressing a critical gap in existing datasets like AVERITEC [Schlichtkrull *et al.*, 2023] and FactCheckQA [Bashlovkina *et al.*, 2023]. The construction involved two key steps: 1) identifying claims likely to retrieve conflicting evidence – particularly those frequently associated with misinformation from untrustworthy sources, and 2) ensuring that retrieved documents for claim verification present conflicting stances.

3.1 Claim Collection

To identify claims likely to retrieve conflicting evidence, we utilized two widely used fact-checking datasets:

- **AVERITEC.** This dataset [Schlichtkrull *et al.*, 2023] contains 4,568 real-world claims fact-checked by 50 organizations, categorized as *Conflicting Evidence/Cherry-picking*, *Not Enough Evidence*, *Refuted*, and *Supported*. We selected claims labeled as *Refuted* or *Supported*, which involve clear factuality.

[4] Dataset available at https://github.com/zoeyyes/CONFACT

- **FactCheckQA.** This dataset [Bashlovkina *et al.*, 2023] includes 20,871 claims annotated as *true*, *false*, or *other*. We focused on claims labeled as *true* or *false*, which provide definitive factuality.

Claims from these datasets were merged[5], covering diverse topics. This process resulted in 3,180 claims: 566 from AVERITEC and 2,614 from FactCheckQA.

3.2 Conflicting Evidence Collection

To facilitate the study of conflicting evidence in fact-checking, we retrieved relevant documents for claim verification. Instead of directly querying Google with the original claims, we transformed each claim into a binary question regarding its veracity using GPT-4[6], following the approach outlined in [Schlichtkrull, 2024; Bashlovkina *et al.*, 2023]. For example, the claim *Nigeria had a population of 45 million at the time of independence"* was converted into the question *Did Nigeria have a population of 45 million at the time of independence?"*. Each question was then submitted as a query on Google, from which we retrieved the top 10 web pages[7]. To ensure reproducibility, the retrieved web pages were archived using the Wayback Machine[8].

3.3 Conflict Evidence Annotation

Next, we annotated the stances of the collected evidence documents using a two-stage process designed to identify conflicting viewpoints.

Stage 1: GPT-4 Annotation. We employed GPT-4 to classify the stance of each document with respect to its corresponding claim as either *supporting* or *refuting*. To enhance robustness, we used three distinct prompt variations: (i) classify the stance based solely on the document URL; (ii) classify the stance using the retrieved webpage content; and (iii) prompt GPT-4 to provide its reasoning prior to making a classification. The specific prompts are detailed in Appendix.

The final stance for each document was determined through majority voting across these three approaches. We defined a claim as exhibiting conflicting evidence if it was associated with documents classified as both *supporting* and *refuting*. Out of 3,180 claims, 611 (17.8%) met this criterion and advanced to the next stage.

Stage 2: Human Annotation. Human annotators subsequently validated the conflicting evidence identified in Stage 1. For each claim, annotators reviewed pairs of documents—one labeled as *supporting* and another as *refuting* by GPT-4. The annotators verified the stances and assessed the credibility of the sources on a 5-point scale (1 = least credible, 5 = most credible). Additionally, they categorized each source into one of the following groups: *Mainstream News*, *Government*, *Non-profit*, *Academic*, *Social Media*, or *Other*. Each document pair was independently reviewed by two annotators, with any disagreements resolved by a third annotator. Detailed annotation guidelines are provided in Appendix.

Split	Labels	# Sources
ModC	125 Yes; 486 No	2469
HumC	51 Yes; 236 No	1418

Table 1: Statistics of the ModC and HumC split of our CONFACT.

3.4 Dataset Analysis

The final CONFACT dataset consists of two splits: *Model Conflicts* (ModC) and *Human Conflicts* (HumC). ModC comprises claims with conflicting evidence identified by GPT-4 during Stage 1. Given that GPT-4 is a powerful closed-source model, this split contains conflicts that may be particularly challenging for most open-source models to resolve. HumC consists of claims where the evidence is conflicting from a human perspective, aiming to assess how effectively fact-checking systems can mitigate human uncertainty when verifying such evidence. The inter-annotator agreement for HumC, as measured by Krippendorff's Alpha, was 0.586—indicating strong agreement while also reflecting the general confusion among annotators when dealing with conflicting documents. Following prior work [Bashlovkina *et al.*, 2023; Schlichtkrull, 2024], we further formulate the claim verification task into a binary question regarding claim veracity, making it more naturally suited for retrieval-augmented LLMs. The binary questions were generated with GPT-4 as discussed in Section 3.2. Claims labeled as *true/supported* correspond to questions with *Yes* as answers, and those labeled as *false/refuted* correspond to questions with *No* as answers. The statistics of CONFACT are provided in Table 1 and an illustration of a data sample from CONFACT is provided in Appendix.

An analysis of document credibility revealed key challenges in assessing source credibility. Annotators frequently overestimated the reliability of *Mainstream News* sources, with 95.8% of these sources rated as credible or neutral. Cross-referencing these ratings with expert annotations from Media Bias / Fact Check (MBFC)[9] showed that 30% of misleading sources were flagged as unreliable by MBFC, while annotators classified 69.8% of these as trustworthy. These findings underscore the challenges of accurately assessing credibility and highlight the importance of addressing conflicting evidence in fact-checking tasks. More details for the distribution of source credibility over source types are available in Appendix.

4 Methodology

In this section, we evaluate retrieval-augmented LLMs on the CONFACT dataset to assess their robustness in fact-checking when confronted with conflicting evidence. We begin by formally defining the task in Section 4.1. Next, in Section 4.2, we describe baseline retrieval-augmented LLMs for

[5]Claims from social media platforms were excluded as they are less findable by search engines.

[6]https://openai.com/index/gpt-4/

[7]Searches and scraping were conducted within a single week (September 12–19, 2024)

[8]https://web.archive.org/

[9]https://mediabiasfactcheck.com/

fact-checking. Finally, Section 4.3 presents strategies for incorporating media source background information at various stages of the RAG pipeline.

4.1 Problem Definition

Given a claim verification question $\mathcal{Q}$ with its relevant N retrieved documents $\{\mathcal{D}_n\}_{n=1}^{N}$ from CONFACT, a retrieval-augmented LLM is expected to generate an answer $\mathcal{A}$ to the question that reflects the veracity of the original claim against available evidence. The system is evaluated on its accuracy in correctly predicting the veracity of claims (i.e., whether $\mathcal{A}$ exactly matches the ground-truth label $\hat{\mathcal{A}}$ for the converted question for claim verification). In addition, we report the *Macro-F1* score as an auxiliary metric to assess performance across classes, particularly given the imbalanced nature of the dataset.

A typical retrieval-augmented fact-checking workflow consists of three main stages: *retrieval*, *ranking*, and *answer generation* [Wang *et al.*, 2024b; Gao *et al.*, 2023], as illustrated in Figure 2(a).

- *Retrieval*: Given a claim verification question $\mathcal{Q}$, a RAG model retrieves relevant documents from an external knowledge base, represented as $\{\mathcal{D}_n\}_{n=1}^{N}$. Retrieved documents were provided on CONFACT to ensure reproducibility, as retrieval is time-varying.
- *Ranking*: Retrieved documents are chunked into short passages, and a ranking function selects the top-K most relevant paragraphs $\{\mathcal{P}_k\}_{k=1}^{K}$ for fact-checking.
- *Answer Generation*: The selected paragraphs $\{\mathcal{P}_k\}_{k=1}^{K}$ are passed to an LLM to generate the final answer $\mathcal{A}$.

4.2 Baseline Retrieval-Augmented LLMs

Baseline retrieval-augmented LLMs adhere to the standard workflow illustrated in Figure 2(a). In this process, the most relevant set of paragraphs $\{\mathcal{P}_k\}_{k=1}^{K}$ are extracted and used as input for answer generation. We evaluate multiple prompting strategies for leveraging these augmented contexts:

- *Direct Answer* (**DirA.**): The K selected paragraphs are provided to the LLM along with the claim verification question, and the model directly generates an answer.
- *Majority Vote* (**MajV.**): The model first predicts answer candidates $\mathcal{A}_k$ for each paragraph $\mathcal{P}_k$. A majority vote is then conducted to select the final answer.
- *Discern and Answer* (**DisA.**): Inspired by [Hong *et al.*, 2024], an explicit instruction is added to filter out misleading passages before generating an answer.
- *Chain-of-Thought* (**CoT**): This strategy prompts the LLM to generate a rationale before predicting the answer [Wei *et al.*, 2022], improving reasoning in multi-step verification tasks.

While these strategies perform well in standard question-answering tasks, they struggle when the retrieved evidence exhibits conflicting viewpoints. For instance, DirA. may conflate misinformation with factual content, MajV. fails if misleading sources outnumber reliable ones, and DisA. depends on the LLM's ability to filter unreliable information, which is not always effective. These limitations motivate the incorporation of media background knowledge.

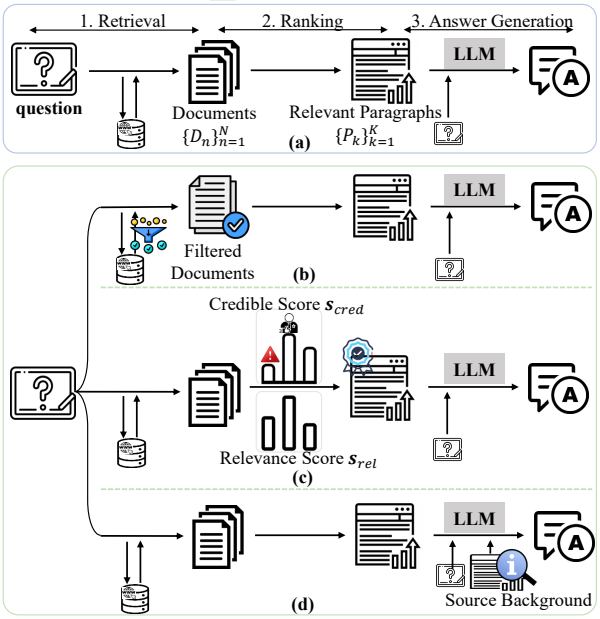

Figure 2: (a) illustrates a general framework of RAG methods involving three stages: retrieval, ranking and answer generation. (b-d) demonstrate our source-aware retrieval-augmented LLMs, incorporating source background information in three stages of the general RAG framework.

4.3 Retrieval-Augmented LLMs with Media Source Backgrounds

To improve fact-checking performance in the presence of conflicting evidence, we propose integrating background information from the source of the media at different stages of the RAG pipeline.

Media Source Background Provider

For each retrieved document, we extract background information about its source. The MBFC website serves as our primary source background provider (GT-MB), offering expert annotations on media bias and factual reliability. If a source is available in MBFC, its credibility rating is retrieved. Otherwise, the background is marked as missing.

To extend coverage beyond MBFC, we introduce a **Hybrid-MB** provider, combining MBFC annotations with an LLM-based background generator [Schlichtkrull, 2024]. The generator first retrieves real-time information about the source's publisher, past credibility ratings, and history of misinformation via Google Search APIs[10]. It then processes this information using an in-context learning approach with a set of pre-defined prompts, generating a credibility summary (denoted as $\mathcal{B}$) that includes factual accuracy, bias, and misinformation history (Refer to Appendix for the designed prompts).

Although the generated source credibility description is comprehensive, it may not be directly applicable at all stages

[10]https://developers.google.com/custom-search/v1/overview

of retrieval-augmented LLMs. Therefore, we further map this description into a credibility score $s_{\text{cred}} \in (0, 1)$ using a prediction model π_θ:

$$s_{\text{cred}} = \pi_\theta(\mathcal{B}). \tag{1}$$

The model is trained on [Baly *et al.*, 2018], which provides labeled credibility supervision. More details about the credible score prediction are provided in Appendix.

Media Background Incorporation

We explore to incorporate source credibility information in three stages of the RAG pipeline:

*1. Source Filtering in Retrieval (**SF**)*: It aims to filter incredible information in the document level. Documents from sources described as *low credible* according to $\mathcal{B}$ are filtered before ranking (Figure 2(b)) (more details in Appendix). The remaining documents are ranked, and the top-K paragraphs are used for answer generation.

*2. Credibility Weighting in Ranking (**CW**)*: Instead of filtering in the document level, credibility scores influence ranking (Figure 2(c)). The final ranking score for a paragraph $\mathcal{P}_m$ is computed as:

$$s_m = s_{\text{rel},m} + \beta * s_{\text{cred},m}, \tag{2}$$

where $s_{\text{rel},m}$ is the relevance score and β balances relevance and credibility. We considered both a soft (CW_{soft}) and a hard (CW_{hard}) setting for leveraging the credible score where CW_{hard} further maps s_{cred} into 0 and 1. Specifically, if s_{cred} is below a threshold γ, it will be mapped to 0, otherwise, 1.

*3. Source Backgrounds Augmentation in Generation (**SBA**)*: Source backgrounds are included at the answer generation stage (Figure 2(d)). We evaluate four strategies:

- $\textbf{SBA}_{\textbf{dir}}$: Concatenates each paragraph with its source background for source-aware paragraphs ($[\mathcal{P}_k, \mathcal{B}_k]$). The K source-aware paragraphs are fed to LLMs for a direct answer.
- $\textbf{SBA}_{\textbf{CoT}}$: Uses CoT prompt with source-aware paragraphs.
- $\textbf{SBA}_{\textbf{exp}}$: Receives source-aware paragraphs and uses explicit instructions to filter unreliable sources.
- $\textbf{SBA}_{\textbf{ens}}$: Uses a two-stage process where candidate answers are generated per paragraph, and conflicts are resolved based on source-aware rationales:

$$\mathcal{A}_k, \mathcal{R}_k = \text{LLM}([\mathcal{P}_k, \mathcal{B}_k], \mathcal{Q}) \tag{3}$$

$$\mathcal{A}^* = \text{LLM}([\mathcal{A}_1, \mathcal{R}_1, \ldots, \mathcal{A}_D, \mathcal{R}_D], \mathcal{Q}) \tag{4}$$

where $\mathcal{A}^*$ is the final answer after considering all rationales.

Refer to Appendix for the designed prompts.

5 Experiments

5.1 Main Experimental Results

We conducted extensive experiments on the ModC and HumC splits of **CONFACT** (for implementation details, see Appendix) to evaluate the performance of retrieval-augmented LLMs in fact-checking scenarios involving conflicting evidence. Our evaluation compares baseline RAG models that do not consider media source backgrounds (Baseline) against models that integrate source credibility data at different stages of the pipeline, using the strategies introduced in Section 4.3 (i.e., Source Filtering (**SF**), Credibility Weighting ($\textbf{CW}_{[\cdot]}$), and Source Background Augmentation ($\textbf{SBA}_{[\cdot]}$)). The experiment results are presented in Table 2. Below, we analyze key findings from our experiments by addressing three research questions.

RQ 1: *How do vanilla retrieval-augmented LLMs perform when confronted with conflicting evidence from sources of varying credibility?*

As shown in the first block of Table 2, vanilla RAG models exhibit difficulties when dealing with conflicting evidence. Their performance is notably limited, as reflected by lower F1 scores, suggesting challenges in correctly classifying the minority class (i.e., claims where the majority of retrieved evidence is misleading). This is primarily due to three key issues. First, hallucination — when presented with conflicting sources, LLMs sometimes generate factually incorrect responses that do not accurately reflect the retrieved evidence. Second, over-reliance on high-frequency responses — the Majority Vote setting biases the system toward the dominant source perspective, often amplifying misinformation if it is overrepresented in retrieval. Third, inability to distinguish misinformation from reliable sources — since vanilla RAG models do not assess source credibility, they treat all retrieved documents as equally valid, leading to incorrect fact-checking outputs. Notably, using GPT-4o in RAG methods (see Appendix) showed no clear advantage over open-source models, highlighting the problem's complexity.

Among the baseline answering strategies, Discern-and-Answer (**DisA.)** and Chain-of-Thought (**CoT**) prompting achieve better results than direct answer generation. This improvement suggests that prompting LLMs to explicitly reason about retrieved content helps mitigate the influence of unreliable sources. However, despite these improvements, the overall accuracy and F1 scores remain suboptimal, highlighting the need for more effective mechanisms to incorporate source credibility into the fact-checking process.

RQ 2: *Does incorporating media source backgrounds improve fact-checking performance in RAG-based LLMs?*

Incorporating media background information into RAG models generally leads to improved performance, although the degree of improvement varies across models. Specifically, LLaMA-3.1 shows a 10% absolute improvement in F1 score, while Mistral achieves a 5% in accuracy improvement when media backgrounds are integrated on the ModC split. Similar improvements are observed on HumC, with the incorporation of source credibility information. These results indicate that providing source credibility cues helps LLMs resolve conflicting evidence more effectively.

However, not all models benefit equally from media backgrounds. Specifically, Qwen-2 exhibits the least improvement, which we attribute to its weaker long-context processing capabilities. The inclusion of source background information significantly increases input length. In models that do not handle extended sequences efficiently, this can dilute relevant context, increase token misalignment, and disrupt self-attention mechanisms, ultimately leading to suboptimal fact-

Set.	Meth.	ModC LLaMA-3.1		ModC Qwen-2		ModC Mistral		HumC LLaMA-3.1		HumC Qwen-2		HumC Mistral	
		Acc.	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**	**Acc.**	**F1**
Bsl.	DirA.	71.36	66.13	78.40	45.96	77.58	70.82	70.03	63.81	77.70	67.83	75.96	67.87
	MajV.	79.87	45.96	79.71	45.90	79.54	45.83	82.93	49.07	82.93	49.07	82.93	49.07
	DisA.	72.18	63.26	78.89	69.54	76.10	70.70	69.69	57.87	80.14	70.04	77.35	71.16
	CoT	77.58	68.50	72.34	67.74	75.46	71.52	75.96	65.50	72.47	64.32	73.87	67.99
GT	SF	71.52	66.28	78.40	69.87	77.74	70.98	69.34	62.97	77.70	67.83	76.31	68.19
	CW_{soft}	67.76	62.75	75.61	65.57	75.12	67.82	67.25	60.45	73.87	61.65	73.52	64.10
	CW_{hard}	68.09	63.62	74.30	64.59	74.47	67.74	68.99	62.93	73.17	61.50	74.22	66.01
	SBA_{dir}	73.16	67.96	79.21	70.51	79.87	73.06	72.82	66.46	78.40	68.11	78.40	69.80
	SBA_{CoT}	78.07	68.32	71.85	67.83	76.92	73.94	78.40	66.07	72.47	67.75	77.00	72.44
	SBA_{exp}	74.47	64.57	80.03	70.13	73.65	68.98	71.40	58.31	80.49	67.44	75.26	69.20
	SBA_{ens}	76.76	68.35	67.92	64.39	66.61	64.53	75.61	65.19	67.94	63.13	68.29	64.79
Hyb.	SF	67.10	62.30	78.07	68.96	75.45	68.62	64.46	58.72	77.70	65.65	74.91	66.31
	CW_{soft}	70.05	65.00	77.41	68.35	77.25	70.02	70.73	64.15	77.00	66.05	77.35	67.85
	CW_{hard}	70.38	65.49	77.74	68.96	76.92	69.72	70.38	64.36	77.35	67.14	75.26	66.30
	SBA_{direct}	74.96	69.30	79.05	70.20	80.03	73.46	74.91	68.06	78.75	68.06	78.75	70.76
	SBA_{CoT}	75.29	65.78	73.00	68.69	75.29	71.90	73.87	61.20	74.22	68.26	75.96	71.14
	SBA_{exp}	76.10	64.53	80.69	70.50	72.83	70.50	75.26	60.88	82.93	70.56	74.91	71.47
	SBA_{ens}	76.76	67.75	66.78	63.25	67.27	63.81	74.91	63.39	64.11	58.01	66.55	60.07

Table 2: Performance of retrieval-augmented LLMs on the **ModC** and **HumC** splits of our CONFACT dataset. *Baseline* (**Bsl.**) denotes models without incorporating source backgrounds.*GT-MB* (**GT**) represents models that only consider incorporating source backgrounds with ground-truth human annotations. *Hybrid-MB* (**Hyb.**) demonstrates models incorporated with both human-annotated source backgrounds as well as automatically generated media backgrounds. The best results (the highest summation of Acc. and F1) are underlined.

checking performance. This finding suggests that as LLM architectures improve in handling long inputs, the benefits of integrating source-aware fact-checking will likely become more pronounced.

RQ 3: *What is the most effective strategy for incorporating media source backgrounds into retrieval-augmented LLMs?*

Different strategies for integrating media backgrounds show distinct patterns of performance across RAG models. Our results indicate that the most effective approach is to incorporate media backgrounds at the answer generation stage, combined with a structured reasoning strategy such as **CoT** prompting or explicit instructions to discern unreliable source.

In contrast, strategies that introduce media backgrounds in earlier stages-such as retrieval or ranking-are less effective. This is likely due to information loss when converting detailed textual source descriptions into a single credibility level or a credibility score. The credibility score predictor, despite being trained on expert-annotated data, does not always provide precise mappings between background descriptions and factual reliability, leading to potential misclassifications.

Furthermore, credibility-aware ranking strategies ($\mathbf{CW_{soft}}$ and $\mathbf{CW_{hard}}$) sometimes degrade performance. This occurs because credibility and relevance are not always aligned—highly credible sources may not contain the most pertinent evidence for verifying a claim. Additionally, credibility-based filtering can risk removing crucial counter-evidence. Fact-checking often requires evaluating misleading claims in context, and aggressively filtering out sources deemed unreliable may leave models without the necessary contrastive information to identify misinformation. As a result, ranking methods that overly rely on credibility scores can paradoxically reduce fact-checking accuracy by limiting the model's ability to reason over conflicting viewpoints.

Comparing GT-MB (which uses expert-verified MBFC credibility labels) and Hybrid-MB (which estimates credibility for missing sources using LLM-based retrieval), we do not observer obvious superiority of Hybrid-MB. This indicates that current source credibility estimation methods remain limited, which could add noise to source credibility aware RAG methods. Manually curated credibility assessments are still more reliable than automated credibility prediction. Detailed error analysis is provided in Appendix.

Summary of Findings. Our results demonstrate that retrieval-augmented LLMs struggle with conflicting evidence when source credibility is not explicitly considered. Integrating media backgrounds improves performance, but the effectiveness of this approach depends on how and where the information is introduced within the pipeline. The most effective strategy is incorporating background information at the answer generation stage, where structured reasoning techniques such as Chain-of-Thought prompting or explicit instructions to discern unreliable source to resolve conflicting claims more effectively. In contrast, relying solely on credibility-aware filtering or ranking may inadvertently introduce biases or remove crucial context needed for fact-checking.

Our findings also reveal a fundamental trade-off between using expert-verified credibility data (GT-MB) and automated credibility estimation (Hybrid-MB). While expert annotations provide higher reliability, automated credibility inference allows for broader source coverage and scalability. Improving the accuracy of LLM-based credibility prediction remains a key open challenge for future research. These insights contribute to the broader field of AI-driven fact-checking by demonstrating both the potential and limita-

Top-K	Chk.	Meth.	LLaMA-3.1		Qwen-2	
			Acc.	F1	Acc.	F1
Top-10	Para.	SBA_{dir}	71.78	65.04	79.09	69.76
		SBA_{CoT}	76.66	66.02	57.14	67.34
		SBA_{exp}	74,22	63.06	65.16	67.86
		SBA_{ens}	74.56	63.62	50.52	48.43
Top-5	Sent.	SBA_{dir}	62.37	57.52	72.82	62.28
		SBA_{CoT}	67.94	58.84	61.32	62.28
		SBA_{exp}	67.25	54.71	75.12	64.18
		SBA_{ens}	68.99	59.76	54.01	53.90

Table 3: Ablation results when using top-10 pieces of augmented context paragraph (para.) and top-5 sentence-level (sent.) chunking strategy.

tions of leveraging source credibility to enhance retrieval-augmented generation for misinformation detection. Limitations are discussed in Appendix.

5.2 Ablation Studies

To further understand the impact of media source backgrounds on fact-checking with conflicting evidence, we conduct ablation studies focusing on GT-MB model on HumC, as they avoid noise from automated source estimation (Hybrid-MB) and HumC is more challenging as shown in Section 5.1. Here, we consider the most powerful way (i.e., in the answer generation stage) to incorporate source credibility information.

Effect of the Number of Augmented Paragraphs. We assess whether increasing the number of retrieved paragraphs improves fact-checking performance by expanding the evidence set from 5 (Table 2) to 10 documents (the first block in Table 3). Surprisingly, this does not enhance accuracy as models may struggle with long inputs as well as be distracted from irrelevant information.

The main reasons are twofold: (1) Increasing the number of retrieved documents introduces lower-relevance evidence, which makes it harder for the model to discern factual correctness. (2) Longer input sequences overwhelm LLM attention mechanisms, leading to poorer factual reasoning. These findings suggest that retrieving fewer but more relevant documents is more effective than increasing retrieval breadth when dealing with conflicting claims.

Impact of Chunking Strategies. We compare paragraph-level (Table 2) vs. sentence-level chunking (the second block in Table 3) for retrieved evidence in the fact-checking pipeline. Paragraph-level chunking consistently outperforms sentence-level chunking, as fragmented sentences often lack sufficient context to resolve factual disputes. However, longer paragraph inputs increase computational overhead.

A potential solution is de-contextualization methods, where sentences are supplemented with surrounding context before being processed by LLMs. Future work could explore such strategies to maintain high-context resolution while minimizing input length constraints.

5.3 Human Evaluation

Beyond the quantitative analysis in Section 5.1, we conduct a qualitative study to assess human fact-checking performance under conflicting evidence. We select 20 fact-checking claims from CONFACT and recruit four NLP researchers as human evaluators. Each human evaluator evaluates 10 claims across three settings, mirroring Section 5.1: (1) without any source background, (2) with curated media backgrounds from MBFC (GT), and (3) with both MBFC-curated and automatically generated source backgrounds (Hybird). The accuracy of human evaluations is summarized in Table 4.

	w/o Background	GT	Hybrid
Acc.	49.45	**50.34**	48.47

Table 4: Human performance on CONFACT without background information, provided with GT background information, and hybrid background information.

Our findings indicate that humans often respond with "unsure" when faced with conflicting evidence, mirroring model performance: while GT media backgrounds boost accuracy, hybrid sources (including AI-generated backgrounds) tend to introduce noise and mislead evaluators. This suggests that unreliable or AI-generated context can impair judgment rather than enhance it.

These results have critical implications for real-world fact-checking organizations. Fact-checkers must adopt rigorous source verification methods to mitigate misinformation risks, and automated tools should prioritize high-fidelity data curation over broad retrieval to reduce misleading noise. Moreover, AI-generated evidence should be treated as assistive rather than authoritative, with human oversight ensuring effective verification of conflicting claims.

Overall, this human evaluation highlights the complexity of fact-checking amid conflicting evidence, reinforcing the need for high-quality evidence retrieval and robust verification mechanisms in both human and automated fact-checking systems.

6 Conclusion

This study presents a systematic evaluation of RAG models in fact-checking scenarios involving conflicting evidence—a critical yet underexplored challenge. To support this, we introduce the CONFACT dataset, which pairs fact-checking claims with contradictory information from sources of varying credibility. Our analysis indicates that existing RAG models struggle when faced with conflicting evidence, often ascribing undue reliability to less credible sources.

To address this issue, we integrate background information from the media sources into the RAG pipelines. Our findings reveal that incorporating source credibility signals during answer generation significantly enhances performance by reducing the models' susceptibility to misinformation. However, challenges remain, particularly in accurately assessing source credibility and mitigating biases in evidence retrieval. These findings highlight the need for automated fact-checking systems to go beyond simple retrieval and ensure rigorous source validation. AI-assisted verification should complement, not replace, human expertise. Future work should refine credibility assessments, enhance evidence ranking, and improve reasoning under uncertainty.

Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research/project is supported by the Ministry of Education, Singapore, under its SUTD-SMU Joint Grant Call, if applicable).

Contribution Statement

Ziyu Ge and Yuhao Wu contributed equally for the paper.

References

[Amplayo *et al.*, 2023] Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. Query refinement prompts for closed-book long-form QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL*, pages 7997–8012, 2023.

[Baly *et al.*, 2018] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, 2018.

[Baly *et al.*, 2020] Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James R. Glass, and Preslav Nakov. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3364–3374, 2020.

[Bashlovkina *et al.*, 2023] Vasilisa Bashlovkina, Zhaobin Kuang, Riley Matthews, Edward Clifford, Yennie Jun, William W. Cohen, and Simon Baumgartner. Trusted source alignment in large language models. *CoRR*, abs/2311.06697, 2023.

[Chen and Shu, 2024] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR*, 2024.

[Chen *et al.*, 2021] Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, pages 4472–4485, 2021.

[Chen *et al.*, 2024] Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 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*, pages 3569–3587, 2024.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023.

[Guo *et al.*, 2022] Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206, 2022.

[Guu *et al.*, 2020] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[Hong *et al.*, 2024] Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In *Findings of the Association for Computational Linguistics: NAACL*, pages 2474–2495, 2024.

[Hounsel *et al.*, 2020] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan R. Mayer. Identifying disinformation websites using infrastructure features. In *10th USENIX Workshop on Free and Open Communications on the Internet, FOCI*, 2020.

[Lee *et al.*, 2024] Yoonsang Lee, Xi Ye, and Eunsol Choi. Ambigdocs: Reasoning across documents on different entities under the same name. *CoRR*, abs/2404.12447, 2024.

[Lewis *et al.*, 2020] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[Li *et al.*, 2023] Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. A survey of large language models attribution. *CoRR*, abs/2311.03731, 2023.

[Min *et al.*, 2020] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 5783–5797, 2020.

[Mukherjee and Weikum, 2015] Subhabrata Mukherjee and Gerhard Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM*, pages 353–362, 2015.

[Nakov *et al.*, 2021] Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto

Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4551–4558, 2021.

[Pan *et al.*, 2023a] Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. Qacheck: A demonstration system for question-guided multi-hop fact-checking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 264–273, 2023.

[Pan *et al.*, 2023b] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 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*, pages 6981–7004, 2023.

[Popat *et al.*, 2016] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM*, pages 2173–2178, 2016.

[Popat *et al.*, 2017] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1003–1012, 2017.

[Ram *et al.*, 2023] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331, 2023.

[Schlichtkrull *et al.*, 2023] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*, 2023.

[Schlichtkrull, 2024] Michael Schlichtkrull. Generating media background checks for automated source critical reasoning. *CoRR*, abs/2409.00781, 2024.

[Thorne *et al.*, 2018] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 809–819, 2018.

[Wang *et al.*, 2024a] Lionel Z. Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. Megafake: A theory-driven dataset of fake news generated by large language models. *CoRR*, abs/2408.11871, 2024.

[Wang *et al.*, 2024b] Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 17716–17736, 2024.

[Wang, 2017] William Yang Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 422–426, 2017.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*, 2022.

[Zhang and Gao, 2024] Xuan Zhang and Wei Gao. Reinforcement retrieval leveraging fine-grained feedback for fact checking news claims with black-box LLM. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 13861–13873, 2024.

[Zhang *et al.*, 2019] Yifan Zhang, Giovanni Da San Martino, Alberto Barrón-Cedeño, Salvatore Romeo, Jisun An, Haewoon Kwak, Todor Staykovski, Israa Jaradat, Georgi Karadzhov, Ramy Baly, Kareem Darwish, James R. Glass, and Preslav Nakov. Tanbih: Get to know what you are reading. 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*, pages 223–228, 2019.