BiDeV: Bilateral Defusing Verification for Complex Claim Fact-Checking

Yuxuan Liu[1], Hongda Sun[1], Wenya Guo[2], Xinyan Xiao[3], Cunli Mao[4], Zhengtao Yu[4], Rui Yan[1]*

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Nankai University
[3] Baidu Inc.
[4] Kunming University of Science and Technology
{yuxuanliu, sunhongda98, ruiyan}@ruc.edu.cn, wenyaguo@nankai.edu.cn,
xiaoxinyan@baudi.com, maocunli@163.com, ztyu@hotmail.com

Abstract

Complex claim fact-checking performs a crucial role in disinformation detection. However, existing fact-checking methods struggle with claim vagueness, specifically in effectively handling latent information and complex relations within claims. Moreover, evidence redundancy, where nonessential information complicates the verification process, remains a significant issue. To tackle these limitations, we propose *Bilateral Defusing Verification* (**BiDeV**), a novel fact-checking working-flow framework integrating multiple role-played LLMs to mimic the human-expert fact-checking process. BiDeV consists of two main modules: *Vagueness Defusing* identifies latent information and resolves complex relations to simplify the claim, and *Redundancy Defusing* eliminates redundant content to enhance the evidence quality. Extensive experimental results on two widely used challenging fact-checking benchmarks (Hover and Feverous-s) demonstrate that our BiDeV can achieve the best performance under both gold and open settings. This highlights the effectiveness of BiDeV in handling complex claims and ensuring precise fact-checking[1].

Introduction

Fact-checking is crucial for claim verification by collecting relevant evidence and determining their veracity (Guo, Schlichtkrull, and Vlachos 2022). Disinformation, concealed within plenty of daily news and reports, threatens the cyber environment and social stability (Liu et al. 2024b). Given its critical role in combating disinformation, complex claim verification has attracted considerable attention from both academics and industry professionals (Thorne and Vlachos 2018; Jiang et al. 2020; Pan et al. 2023).

Recent fact-checking approaches can be broadly categorized into two categories: (i): *Specialized Language Model (SLM)-based end-to-end methods* focus on extracting representations of claims and evidence then comparing them in the feature space for verification (Popat et al. 2018; Soleimani, Monz, and Worring 2020). Typically, they employ specific fine-tuned modules to establish correlations between claims and evidence (Kruengkrai, Yamagishi, and
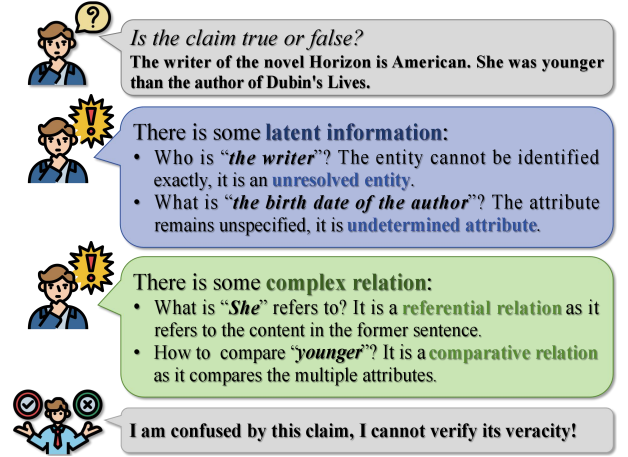

Figure 1: An example illustrating how the claim vagueness impedes the fact-checking process. **Latent information** encompasses unresolved entities and undetermined attributes; **Complex relations** include referential relations and comparative relations.

Wang 2021; Xu et al. 2022; Liao et al. 2023). (ii): *Large Language Model (LLM)-based step-by-step methods* leverage LLMs to conduct questioning or decomposing progressively (Pan et al. 2023; Zhang and Gao 2023; Wang and Shu 2023). These methods benefit from the advanced semantic understanding and reasoning capabilities of LLMs, enabling more nuanced and thorough fact-checking processes.

Despite some promising advancements, several challenges persist in current fact-checking methods, particularly concerning *claim vagueness* and *evidence redundancy*. Claim vagueness poses a primary obstacle in the fact-checking process. Figure 1 illustrates an example of obstacles due to the claim vagueness. In terms of content and correlation, claim vagueness can be categorized into two primary types: (i) *Latent information* encompasses unresolved entities that cannot be identified explicitly and undetermined attributes that remain unspecified. (ii) *Complex relations* include referential relations, where pronouns reference entities within the claim, and comparative relations, which compare multiple attributes. Addressing these aspects is crucial for

---

[1]Code: https://github.com/EthanLeo-LYX/BiDeV

accurately clarifying claims, yet previous approaches often fall short in comprehensively handling these nuances, leading to inadequate verification performance (Liao et al. 2023; Rani et al. 2023). The quality of evidence is essential for claim verification. However, original documents often contain extensive, irrelevant details. This redundancy complicates fact-checking, as current methods overly rely on the evidence and fail to effectively filter out unnecessary information, leading to increased complexity and distraction during the fact-checking process (Zou, Zhang, and Zhao 2023; Zhang and Gao 2023).

To address these challenges, we aim to improve the complex claim fact-checking from two aspects: (i) *claim simplification* identifies the latent information and resolves the complex relations to simplify the claim; (ii) *evidence selection* retains the pertinent evidence and exclude the redundant content. To this end, we propose **Bi**lateral **De**fusing **V**erification (**BiDeV**), a novel complex claim fact-checking working-flow framework that integrates multiple role-played LLMs to imitate the human-expert fact-checking process. To effectively tackle claim vagueness and evidence redundancy, BiDeV incorporates two dedicated modules: (i) **Vagueness Defusing (VD)** formulates claim simplification into two stages: *perceive-then-rewrite* iteratively identifies latent information in the claim, generates corresponding queries for explicit background information, and rewrites the claim for clarity; *decompose-then-check* decomposes the simplified claim, resolves the complex relations, and verifies each sub-claim step by step; (ii) **Redundancy Defusing (RD)** evaluates and filters evidence based on the relevance to specific queries, thus obtaining more precise and pertinent evidence. The VD module aims to simplify claims, reducing the complexity of the fact-checking process by eliminating vagueness. Meanwhile, the RD module enhances the evidence quality by excluding irrelevant content, thus minimizing distractions during verification.

We conduct comprehensive experiments on widely used challenging complex claim fact-checking benchmarks: Hover (Jiang et al. 2020) and Feverous-s (Pan et al. 2023). Experimental results show that BiDeV achieves the best performance, improving Macro-F1 by 3.88% in both annotated evidence (gold) and retrieved evidence (open) settings. This indicates the effectiveness of the proposed VD and RD modules. Also, BiDeV exhibits remarkable improvements on more complex claims, highlighting its competitive generalization ability in handling intricate scenarios.

Overall, our contributions can be summarized as follows:

• We propose BiDeV, a novel fact-checking working-flow framework integrating LLMs to eliminate the vague information in the claim and the noisy redundancy in the evidence, which imitates the fact-checking process of the human experts.

• We introduce the vagueness defusing module formulated as a two-stage process fact-checking a complex claim through perceive-then-rewrite and decompose-then-check. This module concentrates on ascertaining latent information and resolving complex relations, contributing to reducing the complexity of fact-checking complex claims.

• We present the redundancy defusing module to filter out irrelevant information leading to more effective and pertinent evidence in sub-claim verification. Extensive experimental results demonstrate that BiDeV greatly advances the performance in complex claim fact-checking.

Related Work

Complex claim fact-checking aims to identify factual conflicts existing between the claim and the given evidence, which serves as a pivotal technique to address fake news and rumor detection (Liu et al. 2024a).

Previous works can be categorized as *SLM-based end-to-end methods*, which focus on obtaining more effective representations of claims and evidences to conduct verification by comparing them in the feature space (Popat et al. 2018; Ma et al. 2019). Utilizing specific models pre-trained or fine-tuned on some NLI datasets allows them to outperform traditional methods on fact-checking (Kruengkrai, Yamagishi, and Wang 2021; He, Gao, and Chen 2022; Wadden et al. 2022). Moreover, designing some specific modules to correlate the claim and evidence is necessary to achieve more precise verification (Xu et al. 2022; Liao et al. 2023).

As large language models (LLMs) have demonstrated advances in reasoning (Wei et al. 2022; Wang et al. 2022; Sun et al. 2024c), various LLM roles have achieved success in different fields (Sun et al. 2024b,a; Liu et al. 2025). Recent works instruct LLMs to think step-by-step to gradually conduct fact-checking, such as iteratively questioning (Press et al. 2023) and program-guided reasoning (Pan et al. 2023). Some approaches split the complex claim into several simple sub-claims, which reduce the difficulty of verifying each sub-claim (Zhang and Gao 2023; Wang and Shu 2023).

However, previous works have not adequately addressed vague information in the claim and noisy redundancy in the evidence, which limits their performance. To address these issues, we propose BiDeV, which imitates the verification process of human experts, to achieve accurate complex claim fact-checking through more effective claim simplification and evidence selection.

Bilateral Defusing Verification

Task Formulation

The complex claim fact-checking task places a central emphasis on verifying the veracity of the claim based on the pertinent evidence. Specifically, given a claim C, an evidence source S, a fact-checking model M concentrates on predicting the veracity label Y using the evidence from S.

$$Y = M(C, S), \quad Y \in [\text{Support}, \text{Refute}] \quad (1)$$

We address that complex claim fact-checking focuses on claim simplification and evidence selection, thus we formulate our working-flow framework as two main modules: **Vagueness Defusing** for claims and **Redundancy Defusing** for evidence. The overview of our BiDeV is shown in Figure 2. In the subsequent sections, we will introduce how to integrate LLMs to eliminate the vagueness in the claim and the redundancy in the evidence.
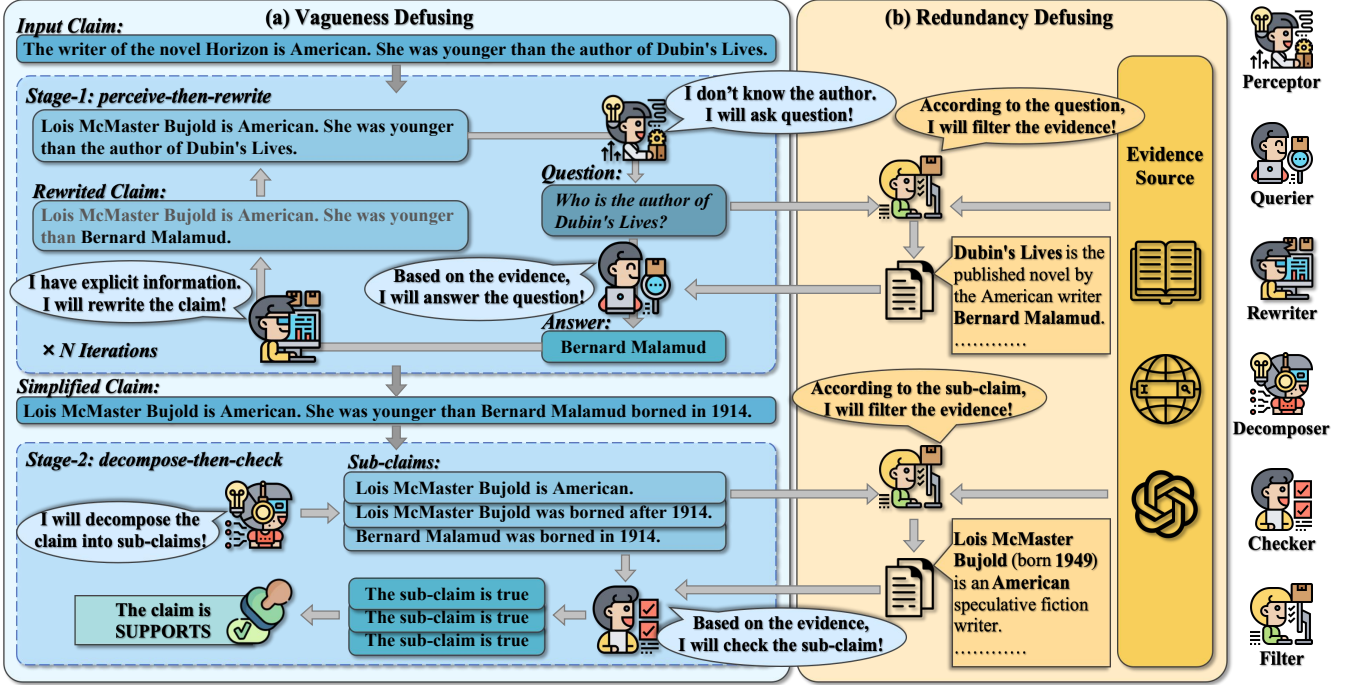
Figure 2: The overview of our BiDeV. Two main modules for Bilateral Defusing Verification: (a) **Vagueness Defusing** for input claim. *Perceive-then-rewrite* stage simplifies the claim iteratively: the perceptor perceives questions about latent information, the querier provides explicit knowledge to the question and the rewriter rewrites the latent information in the claim with the explicit knowledge. *Decompose-then-check* stage verifies the claim: the decomposer splits several sub-claims and the checker verifies the sub-claims. (b) **Redundancy Defusing** for evidence. The evidence extracted from the source is refined by the filter.

Vagueness Defusing

As shown in Figure 1, complex claims often contain two types of vagueness: latent information and complex relations. These elements increase the complexity of fact-checking. When human experts face such complex claims, they first query for undetermined information to obtain explicit background knowledge. Then, they analyze and reconstruct the claim based on the collected background knowledge to eliminate this undetermined information, unravel the complex internal correlations to split several sub-claims, and finally verify the sub-claims to derive the ultimate result (Nakov et al. 2021; Das et al. 2023; Pan et al. 2023). To imitate the iterative process of human experts, we divide the vagueness defusing process into two stages: *perceive-then-rewrite* for latent information and *decompose-then-check* for complex relations.

Stage-1: Perceive-then-Rewrite. Latent information can be classified into two categories: unresolved entities and undetermined attributes. In the example shown in Figure 1, *"the writer"* is an unresolved entity since its reference is not specified within the claim; *"the birth date"* is an undetermined attribute as it is not mentioned in the claim. To defuse the latent information, we instruct LLMs to implement an iterative and collaborative process involving three roles: the perceptor, querier, and rewriter. This process transforms the initial complex claim C into the simplified claim C^* step by step. The details of their working order are discussed below.

• **Perceptor** (M_p) is performed by the LLM through a step-by-step thinking process to perceive latent information in the following standard: (1) An entity is considered to be unresolved if the entity it refers to cannot be found in the claim; (2) An attribute is considered to be undetermined if the attribute of the subject is not mentioned in the claim. Specifically, given the rewritten claim c_{i-1} in the $(i-1)_{th}$ iteration, the perceptor is responsible for accurately identifying both types of latent information and generating targeted question q_i for explicit background knowledge:

$$q_i = M_p(c_{i-1}) \qquad (2)$$

• **Querier** (M_q) is responsible for answering the question generated by the perceptor for precise and explicit content of latent information. Given the question q_i generated by M_p, we instruct the LLM to comprehend and integrate pertinent information within the evidence e_i^* extracted and refined from evidence source S, then generate a precise and dependable answer a_i:

$$a_i = M_q(q_i, e_i^*) \qquad (3)$$

• **Rewriter** (M_r) is essential at this stage as it integrates explicit background knowledge and simplifies the statement of the claim. Since the claim may contain complex internal correlations, merely using these QA pairs as supplementary evidence is insufficient for verification. Given the question q_i in the i_{th} iteration, the rewriter first finds the direct counterparts and the indirect relevance in the claim c_{i-1}, then

rewrites them using the answer a_i. The rewriting process can be formulated as:

$$c_i = M_r(c_{i-1}, q_i, a_i) \qquad (4)$$

Stage-2: Decompose-then-Check. After the perceive-then-rewrite stage, the simplified claim C^* has effectively reduced the latent information but may still contain some complex relations: referential relation and comparative relation. As shown in Figure 1, *"She"* is a referential relation as it refers to *"the writer"* in the former sentence; *"younger"* is a comparative relation as it compares the birth date of *"She"* and *"the author"*. To further clarify claims, we employ a decomposer to disentangle these complex relations and a checker to perform more detailed verification.

• **Decomposer** (M_d) is performed by the LLM to resolve the complex relations: it replaces referential relations with explicit entities in the claim and splits comparative relations using determined attributes. Then the complex claim is decomposed into a series of brief declarative sub-claims with simple logic and unitary content. Given the simplified claim C^*, the process of decomposing sub-claims sc is given by:

$$sc = M_d(C^*) \qquad (5)$$

• **Checker** (M_c) conducts the final step of fact-checking to verify each sub-claims and conclude the veracity result of the entire claim. Since the claim and evidence may describe the same facts in different ways, we guide the LLM to comprehensively understand and extract valuable insights from evidence, then integrate and match them with the claim, and finally produce a dependable result after meticulous reasoning. With the relevant evidence e_j^*, we obtain the verification result y_j of the sub-claims sc_j, and ultimately conclude the predicted veracity label Y of the entire claim:

$$Y = \bigcap_j^{|sc|} y_j, \quad y_j = M_c(sc_j, e_j^*) \qquad (6)$$

Redundancy Defusing

When answering questions and verifying claims, human experts first extract potential evidence and then select pertinent paragraphs providing precise and credible information. Hence, we emulate this process by initially extracting coarse-grained relevant evidence from the evidence source S. For the gold setting, we directly use the evidence from annotated with the gold labels in the dataset. For the open setting, we retrieve evidence from external knowledge bases (*e.g.,* Wikipedia). However, the initially extracted evidence often contains redundant and noisy information, which can confuse the querier and checker. Therefore, we filter out the irrelevant information through step-by-step thinking.

• **Filter** (M_f) firstly segments the initially extracted evidence into multiple paragraphs and then evaluates whether each paragraph is relevant to the question or the sub-claim, which involves not only directly relevant content but also potentially contributed information. The irrelevant paragraphs are eliminated to get the most imperative and effective evidence. As an example of answering a question q_i, given the extracted evidence e_i, we obtain the filtered evidence e_i^* by:

$$e_i^* = M_f(e_i, q_i) \qquad (7)$$

Experiments

Experimental Setup

Datasets. There are two widely used and challenging datasets to evaluate the fact-checking performance of baselines and our BiDeV: (i) **Hover** (Jiang et al. 2020) and (ii) **Feverous-s** (Pan et al. 2023). Both of the datasets need to verify the given claim with multiple evidences through multi-step reasoning.

Baselines. To demonstrate the effectiveness of our method, we compare BiDeV with the following four types of baselines: (i) *Pre-trained methods*: BERT-FC (Soleimani, Monz, and Worring 2020) and LisT5 (Jiang, Pradeep, and Lin 2021). (ii) *Fine-tuned methods*: RoBERTa-NLI (Nie et al. 2020), DeBERTaV3-NLI (He, Gao, and Chen 2022) and MULTIVERS (Wadden et al. 2022). (iii) *LLM-ICL methods*: FLAN-T5 (Chung et al. 2024) and Codex (Chen et al. 2021). (iv) *LLM-reason methods*: Hiss (Zhang and Gao 2023), FOLK (Wang and Shu 2023), ProgramFC (Pan et al. 2023) and FactcheckGPT (Wang et al. 2024).

Evaluation Metrics. We use Macro-F1 as metrics in order to better deal with unbalanced proportions between *support* and *refute* samples.

Implementation Details. In our proposed method, we use `gpt-3.5-turbo` as the base model of Perceptor, Rewriter, Decomposer, and Filter by accessing to OpenAI API with few-shot demonstrations. For a fair comparison, we leverage Flan-T5-XL (3B) as the Querier and Checker without additional fine-tuning. In the vagueness defusing, we iteratively perceive-then-rewrite for 3 rounds. To evaluate in the open setting, we use BM25 (Robertson, Zaragoza et al. 2009) to retrieve top-K (K=10) evidence documents.

Overall Performance

We evaluate BiDeV and the compared baselines on two challenging benchmarks under two settings: annotated evidence as gold-setting and retrieved evidence as open-setting. The overall performance is shown in Table 1. The experimental results demonstrate the following conclusions.

• *BiDeV achieves the best performance.* Our BiDeV achieves appealing performance improvement against 11 baselines from 4 categories. Specifically, BiDeV outperforms fine-tuned baselines by 10.69% (gold) and 15.27% (open) on average without training. Compared to both LLM-based baselines, BiDeV also obtains 6.22% performance improvement. The experiment results demonstrate that our proposed BiDeV could achieve outstanding performance gains.

• *BiDeV improves on more complex claims.* Although DeBERTaV3-NLI could be competitive with BiDeV on easier 2-hop claims, its performance drops extremely as the complexity increases, and BiDeV surpasses it by 5.33% and 12.31% on 3-hop and 4-hop claims. Overall, BiDeV achieves improvement by 10.86%@2-hop, 11.72%@3-hop, and 17.72%@4-hop under gold-setting, which indicates that BiDeV performs more effectively on complex claims.

• *Integrating perceiving, rewriting, and decomposing is effective.* Compared with decomposition-based Hiss, question-based FOLK, and program-guided ProgramFC,

Methods	Hover(hop2)		Hover(hop3)		Hover(hop4)		Feverous-s	
	Gold	Open	Gold	Open	Gold	Open	Gold	Open
Bert-FC* (Soleimani, Monz, and Worring 2020)	53.41	50.68	50.91	49.86	50.86	48.57	74.71	51.67
LisT5* (Jiang, Pradeep, and Lin 2021)	56.15	52.56	53.76	51.89	51.67	50.46	77.88	54.15
RoBERTa-NLI* (Nie et al. 2020)	74.62	63.62	62.23	53.99	57.98	52.41	88.28	57.81
DeBERTaV3-NLI* (He, Gao, and Chen 2022)	77.22	68.72	65.98	60.76	60.49	56.01	91.98	58.81
MULTIVERS* (Wadden et al. 2022)	68.86	60.17	59.87	52.55	55.67	51.86	86.03	56.61
Codex* (Chen et al. 2021)	70.63	65.07	66.46	56.63	63.49	57.27	89.77	62.58
FLAN-T5* (Chung et al. 2024)	73.69	69.02	65.66	60.23	58.08	55.42	90.81	63.73
HiSS† (Zhang and Gao 2023)	73.06	66.25	65.14	58.56	64.67	57.64	89.26	65.99
FOLK† (Wang and Shu 2023)	73.24	67.29	65.84	58.61	64.73	58.79	89.52	66.89
ProgramFC† (Pan et al. 2023)	74.59	69.89	66.75	61.21	65.00	58.21	91.23	67.22
Factcheck-GPT† (Wang et al. 2024)	74.88	70.25	66.32	60.11	66.62	59.25	91.39	67.24
BiDeV	**77.59**	**73.44**	**69.91**	**63.62**	**70.63**	**60.41**	**92.39**	**69.01**

Table 1: Macro-F1 scores of BiDeV and baselines on Hover and Feverous-s under both gold and open settings. Compared baselines include: (i) *Pre-trained methods*; (ii) *Fine-tuned methods*; (iii) *LLM-ICL methods*; and (iv) *LLM-reason methods*. Bold numbers indicate significant improvements ($p < 0.05$) based on 10 rounds of bootstrapping sampling. Results with * are quoted from (Pan et al. 2023), and results with † are reproduced by `gpt-3.5-turbo` for a fair comparison.
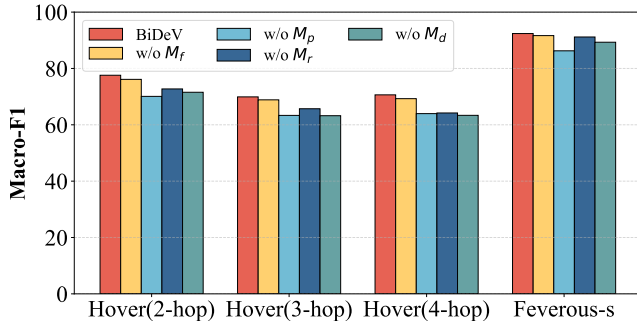

Figure 3: Ablation study on Hover and Feverous-S. M_f: Filter; M_p: Perceptor; M_r: Rewriter; M_d: Decomposer.

BiDeV surpasses them by 5.67% on average, which demonstrates that the integration of perceiving, rewriting, and decomposing could inject explicit background information, resolve intricate correlations, simplify the claim, and reduce the complexity of fact-checking.

Ablation Study

In this section, we eliminate perceptor, rewriter, decomposer, and filter respectively, and explore to what extent these modules have an impact on the complex claim fact-checking. We conducted an ablation study on the gold setting, which is more representative because of its balanced performance. As shown in Figure 3, the perceptor has the most impact, which indicates that generating questions for explicit background information is necessary. The decomposer contributes to the verification as it disentangles a complex claim into several brief sub-claims that are much easier
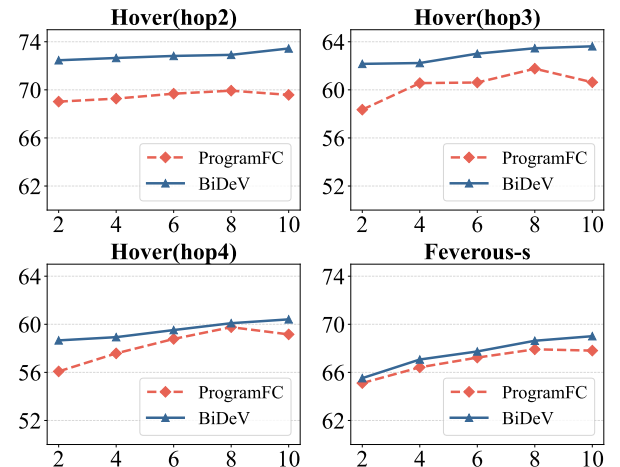

Figure 4: Analysis of Redundancy Defusing under different Top-K retrieved evidence.

to be verified. The rewriter could reduce the complexity of understanding claims as well. These three modules demonstrate that vagueness defusing is effective in simplifying the claim and leading to better fact-checking accuracy. The feasibility of the filter has also confirmed that redundancy defusing could estimate and improve the evidence quality.

Additional Analysis

Analysis of Redundancy Defusing. We conducted comparative experiments with a selected strong baseline ProgramFC to explore the performance in retrieving different numbers of evidence and the results are shown in Figure 4.
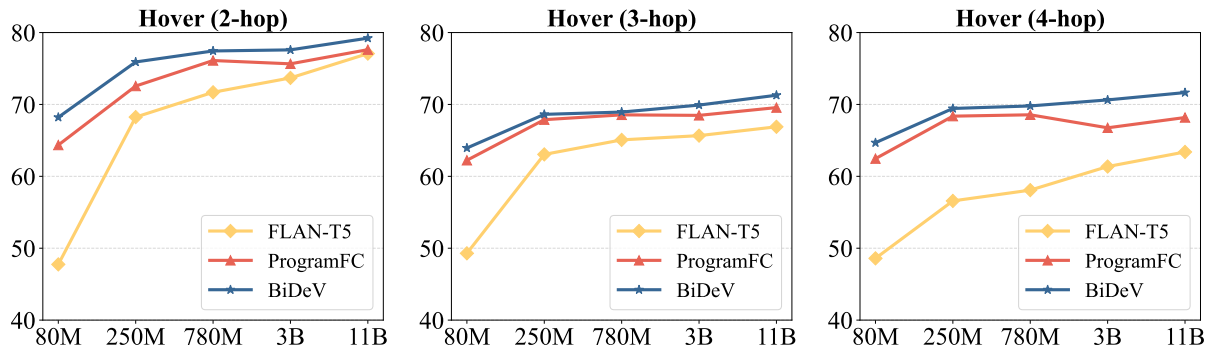
Figure 5: Analysis of different model scales in Querier and Checker: FLAN-T5-small (80M), FLAN-T5-base (250M), FLAN-T5-large (780M), FLAN-T5-XL (3B), FLAN-T5-XXL (11B) on Hover 2-hop, 3-hop, and 4-hop subsets.
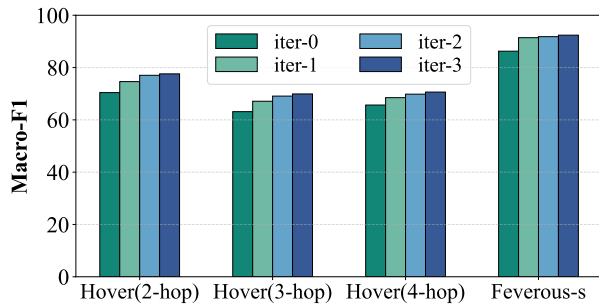

Figure 6: Analysis of different iteration numbers of *perceive-then-rewrite* in Vagueness Defusing.

Strategy	Hover			Feverous-s
	2-hop	**3-hop**	**4-hop**	
Direct	71.55	63.23	63.36	89.31
Naive	73.61	65.34	65.57	90.15
BiDeV	**77.59**	**69.91**	**70.63**	**92.39**
Direct	66.26	59.23	57.17	60.28
Naive	66.85	60.05	57.65	65.41
BiDeV	**73.44**	**63.62**	**60.41**	**69.01**

Table 2: Analysis of different decomposition strategies. Above is the gold setting; Below is the open setting.

Intuitively, more evidence will provide more information, leading to more accurate fact-checking. However, the performance of ProgramFC exhibits an upward-then-downward trend as the number increases, because the gain from useful information is offset by the interference from redundant information when too many evidences are retrieved. Compared to ProgramFC, BiDeV achieves consistent performance improvement as the number of retrieved evidences increases, which demonstrates that redundancy defusing module performs fine-grained filtering from the extracted evidences to obtain more pertinent and effective information.

Analysis of Vagueness Defusing.

• **Iteration of perceive-then-rewrite.** *Perceive-then-rewrite* is an iterative process designed to involve more precise information and simplify the complex claim. We designed experiments to investigate the effect of different numbers of iterations on the verification accuracy. As shown in Figure 6, with the increase in the number of iterations, the accuracy of fact-checking gradually increases and tends to stabilize. The experimental results reveal that it is necessary and effective to constantly rewrite the claim based on queried explicit background knowledge, which eliminates vague information and simplifies the claim. In the trade-off between performance and cost, we finally set the maximum number of iterations to 3 according to the results.

• **Strategies of decomposition.** Decomposition plays an important role in *decompose-then-verify*, thus we explore the

effects of different decomposition strategies: (1) Direct: directly verify the simplified claim; (2) Naive: naively decompose the simplified claim; (3) BiDeV: decompose the simplified claim to resolve complex relations. We conduct evaluation on both gold and open settings, the experimental result is shown in Table 2. Comparison with Direct demonstrates the necessity of the decomposition, and comparison with Naive proves the effectiveness of the complex relation-oriented decomposition in BiDeV.

Analysis of Querier and Checker. In our proposed BiDeV, the accuracy of answering the questions affects the effectiveness of claim rewriting, and the verification of the sub-claim directly influences the overall fact-checking accuracy. Consequently, we scale the base model of Querier and Checker and conduct a comparison on Hover that is more direct to evaluate performance on different complexity of claims (Pan et al. 2023). As shown in Figure 5, our BiDeV allows for better generalization on larger-scale base models. Compared to FLAN-T5, the improvement is more on a smaller base model, 35.22% on 80M parameters, because the reasoning ability is constrained by the model scale. Our bilateral defusing effectively alleviates this issue by simplifying the claim and selecting pertinent evidence. It reveals that BiDeV better eliminates the obstacles to verifying complex claims that we surpass ProgramFC with the sub-task solver of 11B by only using the base model of 250M as Querier and Checker.

Charles Russell directed a 1994 superhero fantasy film. The host of the show VJ Logan won had a part in it.	**Supports**

FOLK

Predicates:
Directed(Charles Russell, a 1994 superhero fantasy film) :::
    Charles Russell directed a 1994 superhero fantasy film
Host(VJ Logan, a show) :::
    VJ Logan host a show
Part(VJ Logan, a 1994 superhero fantasy film) :::
    VJ Logan had a part in the 1994 superhero fantasy film.
Followup Question:
What film did Charles Russell direct in 1994? → _The Mask_
What show did VJ Logan host? → _Unknown!_
Did VJ Logan have a part in the 1994 superhero fantasy film? → _No!_
Predicted Label: Refutes ×

ProgramFC

```
def program():
    answer_1 = Question("Who directed the 1994 superhero fantasy film?")
            Charles Russell
    fact_1 = Verify(f"Charles Russell directed the film {answer_1}.")
            False
    answer_2 = Question("Who won the show VJ Logan?")
            VJ Logan
    fact_2 = Verify(f"{answer_2} had a part in the film.")
            False
    label = Predict(fact_1 and fact_2)
            False
```
Predicted Label: Refutes ×

BiDeV

Perceived question in iter-1:
What is the 1994 superhero fantasy film directed by Charles Russell? → _The Mask_
Rewrited claim in iter-1:
Charles Russell directed **_The Mask_**, a 1994 superhero fantasy film. The host of the show VJ Logan won had a part in **_The Mask_**.

Perceived question in iter-2:
What is the show that VJ Logan won? → _America's Most Smartest Model_
Rewrited claim in iter-2:
Charles Russell directed The Mask, a 1994 superhero fantasy film. The host of **_America's Most Smartest Model_** which VJ Logan won had a part in The Mask.

Perceived question in iter-3:
Who is the host of the show VJ Logan won? → _Ben Stein_
Rewrited claim in iter-3:
Charles Russell directed The Mask, a 1994 superhero fantasy film. **_Ben Stein_**, the host of America's Most Smartest Model which VJ Logan won, had a part in The Mask.

Decomposed Sub-claims:
Charles Russell directed The Mask, a 1994 superhero fantasy film. → _True_
Ben Stein had a part in The Mask. → _True_
Ben Stein was the host of America's Most Smartest Model. → _True_
VJ Logan won America's Most Smartest Model. → _True_

Predicted Label: Supports √

Figure 7: Case Study of selected baselines (FOLK and ProgramFC) and our BiDeV.

Analysis of complex claim comprehension. We also conducted a close-setting experiment that _no_ evidence is available and the model can only achieve better performance by comprehending and simplifying the claim more effectively. The experimental results are shown in Table 3. We surpass FLAN-T5 by 9.62% on average, which indicates that our simplified claim can be checked effectively even by the limited knowledge stored in the parameters of the 3B model. Compared to different reasoning prompt methods of LLM, our BiDeV achieves 5.49% improvement on average. This demonstrates that vagueness defusing contributes to reducing the complexity of fact-checking. Moreover, BiDeV also gains more improvement on more complex claims: 5.76% on 2-hop and 6.39% on 4-hop, which proves that our vagueness defusing module is more effective on complex claims.

Case Study

To present a more intuitive presentation of BiDeV in the fact-checking process, we select FOLK and ProgramFC for comparison. As shown in Figure 7, the vague information in the claim has been eliminated after _perceive-then-rewrite_ stage and the decomposed sub-claims have been verified successfully. However, FOLK generates invalid predicates leading to improper answers to the follow-up questions confused by the complex statement in the claim. Similarly, ProgramFC encounters wrong variable correlation and sub-task function calls. Both FOLK and ProgramFC are close to machine-centric reasoning, which is constrained by complex claims. In contrast, BiDeV imitates the thinking process of human experts achieving more accurate fact-checking.

Methods	**Hover**			**Feverous-s**
	2-hop	**3-hop**	**4-hop**	
FLAN-T5*	48.27	52.11	51.13	55.16
Direct*	56.51	51.75	49.68	60.13
CoT*	50.31	52.32	51.58	54.78
Self-Ask*	51.54	51.47	52.45	56.82
ProgranFC[†]	54.43	54.23	52.74	59.69
BiDeV	**56.73**	**54.82**	**53.66**	**61.12**

Table 3: Analysis of complex claim fact-checking under close-setting. Results with * are quoted from (Pan et al. 2023); Results with [†] are reproduced by `gpt-3.5-turbo`.

Conclusion

We propose Bilateral Defusing Verification (BiDeV) in this paper, a novel framework integrating multiple LLMs to effectively imitate the complex claim fact-checking process of human experts. The vagueness defusing module eliminates latent information and resolves complex correlations, thereby simplifying the claims. The redundancy defusing module filters out irrelevant evidence to provide more pertinent information for verification. Experimental results show that BiDeV advances the best performance on two challenging benchmarks (Hover and Feverous-s). This highlights BiDeV's significant improvements in handling complex claims and offering more intuitive reasoning processes.

Acknowledgements

References

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H. P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; Ray, A.; Puri, R.; Krueger, G.; Petrov, M.; Khlaaf, H.; Sastry, G.; Mishkin, P.; Chan, B.; Gray, S.; Ryder, N.; Pavlov, M.; Power, A.; Kaiser, L.; Bavarian, M.; Winter, C.; Tillet, P.; Such, F. P.; Cummings, D.; Plappert, M.; Chantzis, F.; Barnes, E.; Herbert-Voss, A.; Guss, W. H.; Nichol, A.; Paino, A.; Tezak, N.; Tang, J.; Babuschkin, I.; Balaji, S.; Jain, S.; Saunders, W.; Hesse, C.; Carr, A. N.; Leike, J.; Achiam, J.; Misra, V.; Morikawa, E.; Radford, A.; Knight, M.; Brundage, M.; Murati, M.; Mayer, K.; Welinder, P.; McGrew, B.; Amodei, D.; McCandlish, S.; Sutskever, I.; and Zaremba, W. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70): 1–53.

Das, A.; Liu, H.; Kovatchev, V.; and Lease, M. 2023. The state of human-centered NLP technology for fact-checking. *Information processing & management*, 60(2): 103219.

Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.

He, P.; Gao, J.; and Chen, W. 2022. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations*.

Jiang, K.; Pradeep, R.; and Lin, J. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 402–410.

Jiang, Y.; Bordia, S.; Zhong, Z.; Dognin, C.; Singh, M.; and Bansal, M. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3441–3460.

Kruengkrai, C.; Yamagishi, J.; and Wang, X. 2021. A Multi-Level Attention Model for Evidence-Based Fact Checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2447–2460.

Liao, H.; Peng, J.; Huang, Z.; Zhang, W.; Li, G.; Shu, K.; and Xie, X. 2023. MUSER: A MUlti-Step Evidence Retrieval Enhancement Framework for Fake News Detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4461–4472.

Liu, Y.; Chen, X.; Zhang, X.; Gao, X.; Zhang, J.; and Yan, R. 2024a. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. *arXiv preprint arXiv:2403.09498*.

Liu, Y.; Song, Z.; Zhang, X.; Chen, X.; and Yan, R. 2024b. From a tiny slip to a giant leap: An llm-based simulation for fake news evolution. *arXiv preprint arXiv:2410.19064*.

Liu, Y.; Sun, H.; Liu, W.; Luan, J.; Du, B.; and Yan, R. 2025. MobileSteward: Integrating Multiple App-Oriented Agents with Self-Evolution to Automate Cross-App Instructions. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*.

Ma, J.; Gao, W.; Joty, S.; and Wong, K.-F. 2019. Sentence-Level Evidence Embedding for Claim Verification with Hierarchical Attention Networks. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2561–2571. Florence, Italy: Association for Computational Linguistics.

Nakov, P.; Corney, D.; Hasanain, M.; Alam, F.; Elsayed, T.; Barrón-Cedeño, A.; Papotti, P.; Shaar, S.; and Da San Martino, G. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4885–4901. Online: Association for Computational Linguistics.

Pan, L.; Wu, X.; Lu, X.; Luu, A. T.; Wang, W. Y.; Kan, M.-Y.; and Nakov, P. 2023. Fact-Checking Complex Claims with Program-Guided Reasoning. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6981–7004. Toronto, Canada: Association for Computational Linguistics.

Popat, K.; Mukherjee, S.; Yates, A.; and Weikum, G. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 22–32.

Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N.; and Lewis, M. 2023. Measuring and Narrowing the Compositionality Gap in Language Models. In Bouamor, H.; Pino,

J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5687–5711. Singapore: Association for Computational Linguistics.

Rani, A.; Tonmoy, S. T. I.; Dalal, D.; Gautam, S.; Chakraborty, M.; Chadha, A.; Sheth, A.; and Das, A. 2023. FACTIFY-5WQA: 5W Aspect-based Fact Verification through Question Answering. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10421–10440. Toronto, Canada: Association for Computational Linguistics.

Robertson, S.; Zaragoza, H.; et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4): 333–389.

Soleimani, A.; Monz, C.; and Worring, M. 2020. Bert for evidence retrieval and claim verification. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, 359–366. Springer.

Sun, H.; Lin, H.; Yan, H.; Zhu, C.; Song, Y.; Gao, X.; Shang, S.; and Yan, R. 2024a. Facilitating Multi-Role and Multi-Behavior Collaboration of Large Language Models for Online Job Seeking and Recruiting. *arXiv preprint arXiv:2405.18113*.

Sun, H.; Liu, Y.; Wu, C.; Yan, H.; Tai, C.; Gao, X.; Shang, S.; and Yan, R. 2024b. Harnessing Multi-Role Capabilities of Large Language Models for Open-Domain Question Answering. In *Proceedings of the ACM on Web Conference 2024*, 4372–4382.

Sun, H.; Xu, W.; Liu, W.; Luan, J.; Wang, B.; Shang, S.; Wen, J.-R.; and Yan, R. 2024c. Determlr: Augmenting llm-based logical reasoning from indeterminacy to determinacy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9828–9862.

Thorne, J.; and Vlachos, A. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3346–3359.

Wadden, D.; Lo, K.; Wang, L.; Cohan, A.; Beltagy, I.; and Hajishirzi, H. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 61–76.

Wang, H.; and Shu, K. 2023. Explainable Claim Verification via Knowledge-Grounded Reasoning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6288–6304.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wang, Y.; Reddy, R. G.; Mujahid, Z. M.; Arora, A.; Rubashevskii, A.; Geng, J.; Afzal, O. M.; Pan, L.; Borenstein, N.; Pillai, A.; Augenstein, I.; Gurevych, I.; and Nakov, P. 2024.

Factcheck-Bench: Fine-Grained Evaluation Benchmark for Automatic Fact-checkers. arXiv:2311.09000.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Xu, W.; Wu, J.; Liu, Q.; Wu, S.; and Wang, L. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*, 2501–2510.

Zhang, X.; and Gao, W. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *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 (Volume 1: Long Papers)*, 996–1011. Nusa Dua, Bali: Association for Computational Linguistics.

Zou, A.; Zhang, Z.; and Zhao, H. 2023. Decker: Double Check with Heterogeneous Knowledge for Commonsense Fact Verification. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 11891–11904. Toronto, Canada: Association for Computational Linguistics.