Debate-to-Consensus: Debate-Based Multi-Perspective Fusion for Claim Verification

Anonymous ACL submission

Abstract

The rapid spread of online misinformation has intensified the need for robust claim verification. Previous methods tend to evaluate claims from a single perspective, which often introduces bias. Although recent studies have begun to incorporate multiple analytical dimensions, they frequently overlook the intricate conflicts and complementarities among them. To address this limitation, we propose Debate-to-Consensus, a debate-inspired multi-perspective claim verification framework designed to model and harness these inter-perspective relationships. Inspired by the real-world fact-checking process, our method evaluates claims on three distinct perspectives: statement plausibility, retrieved-source credibility, and evidence consistency. Rather than treating these analyses as isolated information to concatenate, Debate-to-Consensus uses an debate fusion network that estimates pairwise perspective agreement and interpolates between consensus-oriented aggregation and conflict-oriented cross-attention, eventually forging a unified consensus judgment. Experiments on three benchmark datasets demonstrate that Debate-to-Consensus consistently outperforms PLM-only, LLM-only, and multi-perspective baselines, while ablation studies demonstrate the effectiveness of both the perspective design and the debate-inspired fusion mechanism.

1 Introduction

With the rapid growth of the Internet and social networks, various forms of misinformation have spread quickly online, creating an urgent need for claim verification (Chern et al., 2023; Cheung and Lam, 2023; Liu et al., 2024). This problem has been further exacerbated by recent advances in generative AI, which enable the large-scale production of fluent and seemingly credible misinformation, increasing both the volume and persuasiveness of such content (Hu et al., 2024; Garry et al., 2024). Accurate assessment of claims remains highly dependent on the availability of relevant evidence and careful analysis of the relationships between claims and evidence, making effective verification particularly challenging (Popat et al.; Ma et al., 2024; Schlichtkrull et al., 2023).

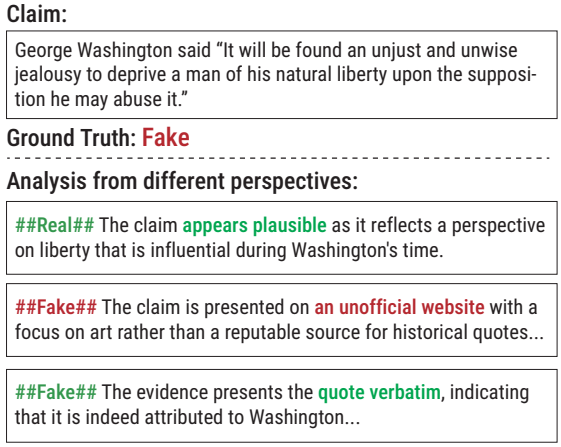

Figure 1: An example that analyzing a fake claim from perspectives with different intermediate judgments, motivating explicit modeling of agreement and conflict among perspectives.

Previous approaches primarily determine the veracity of claims through a singular perspective, typically evaluating the semantic alignment between claims and evidence. These methods focus on joint encoding schemes to capture the compatibility between a claim and its associated references (Yue et al., 2024; Popat et al.).

However, analyzing claims through a singular perspective frequently results in fragmented or mutually inconsistent judgments (Vlachos and Riedel, 2014), as exemplified in Fig. 1. This divergence underscores the inherent insufficiency of isolated perspectives in achieving robust verification.

Recent studies have increasingly shifted toward multi-perspective approaches. Some methods focus on enriching contextual perspectives, using generative models to synthesize enriched back-

ground information for claims, thus alleviating evidence scarcity and uncovering deep-seated textual patterns (Tan et al., 2025; Wu et al., 2024). In addition, other studies involve fine-grained evidence analysis, in which researchers meticulously examine the logical consistency and substance of supporting documents from multiple points of view to determine their veracity (Li et al., 2024; Irnawan et al., 2025; Wang and Shu, 2023).

However, existing multi-perspective approaches typically rely on feature concatenation or late-stage aggregation, which treats each perspective as an isolated signal and seldom account for the underlying inter-relationships between perspectives. Existing designs can use more information for claim verification, but it does not directly distinguish two qualitatively different cases: multiple perspectives independently support the same verdict, or different perspectives expose a conflict that requires closer comparison. Consequently, robust claim verification requires mechanisms that can model both agreement and disagreement among perspectives rather than simply accumulating them (Soprano et al., 2024).

To address these limitations, we propose **Debate-to-Consensus**, a debate-inspired framework that integrates complementary perspectives through debate fusion. Inspired by professional fact-checking workflows, our framework examines each claim through three distinct yet interconnected perspectives: *statement plausibility* assessing the claim's plausibility based on commonsense and logic, *retrieved-source credibility* evaluating the credibility of sources returned by retrieval rather than the original provenance of the claim, and *evidence consistency* verifying whether retrieved evidence supports or contradicts the claim. We then employ a debate-inspired network that allows different perspectives to interact and cross-examine one another, ultimately reaching a consensus on the final judgment.

Experimental results on three benchmark datasets demonstrate that our approach consistently outperforms existing baselines, validating the effectiveness of multi-perspective analysis and highlighting the importance of modeling inter-perspective interactions to achieve more reliable claim verification.

Our contributions include the following.

- We propose a multi-perspective reasoning paradigm that analyzes claims from statement plausibility, retrieved-source credibility, and evidence consistency, enabling a more comprehensive and robust verification process.
- We introduce Debate-to-Consensus, a claim verification framework that generates structured analyzes based on the proposed paradigm and employs an debate-inspired fusion network to model complementary and conflicting relationships among perspectives and reach a unified judgment.
- Experiment results on three benchmark datasets show that Debate-to-Consensus outperforms state-of-the-art methods, and ablation studies verify the value of both multi-perspective analysis and adaptive fusion.

The source code for Debate-to-Consensus is available at anonymous repository.

2 Related Work

2.1 Single-Perspective Claim Analysis

Early approaches primarily analyze claims from an evidence-matching perspective, modeling their semantic alignment with retrieved evidence. These methods encode claims and evidence into shared semantic representations and assess their relationships using specific modules (Popat et al.; Xu et al., 2022; Zheng et al., 2024). Some studies further extend this paradigm by incorporating multi-modal evidence, such as visual information (Wu et al., 2021). Beyond single-granularity representations, Wei et al. (Wei et al., 2025) proposed a dual-granularity architecture that jointly models evidence at different levels, implicitly capturing multiple analytical views within the claim-evidence interaction.

With their strong reasoning capacity, LLMs have been applied to the verification of claims from an evidence-analysis perspective. Researchers employ LLMs to explicitly analyze relationships between claims and evidence, allowing the capture of implicit logic and fine-grained correlations (Tan et al., 2025; Ma et al., 2024). Some studies further analyze claims from a claim reformulation perspective and enrich contextual information by transforming the original formulations of claims. For example, Wu et al. (Wu et al., 2024) generate semantic-flipped and semantic-invariant variants of claims to capture diverse semantic interpretations, while Wang et al. (Wang and Shu,

2023) reformulate claims into first-order logic clauses to facilitate reasoning over evidence.

Despite their strong reasoning capabilities, LLMs are prone to hallucinations and biased judgments. To mitigate these issues, recent studies introduce external information sources into claim analysis using retrieval-augmented pipelines. These approaches iteratively retrieve and assess evidence, using LLMs to determine whether the accumulated information is sufficient to verify the claim (Li et al., 2024; Cheung and Lam, 2023).

2.2 Multi-Perspective Claim Analysis

Verification based on a single perspective is prone to bias, which has motivated recent work to incorporate multiple perspectives to capture diverse signals for more reliable claim verification.

One line of research uses LLMs to simulate public discourse surrounding a claim, providing rich contextual signals. For example, Nan et al. (Nan et al., 2024) prompt LLMs to role-play as users and produce multi-perspective comments, while Wan et al. (Wan et al., 2024) model news reactions to generate additional contextual signals beyond the original claim. These approaches exploit the generative capacity of LLMs to create rich contexts that simulate different perspectives, resulting in more comprehensive claim verification.

In contrast, other methods focus on generating multiple analytical perspectives for the claim itself. For example, LLMs can be guided to analyze claims using common sense reasoning or textual descriptions, each perspective modeled independently in relation to the claim (Hu et al., 2024; Irnawan et al., 2025). Some methods further promote a dialectical verification process by producing adversarial reasoning from opposing viewpoints and concatenate them together to enable the model to reconcile conflicting perspectives (Zheng et al., 2025; Qin and Zhang, 2024). In addition, certain approaches decompose claims into multiple sub-claims, retrieve evidence for each component, concatenate retrieved results and send to LLM to form a final decision (Liu et al., 2024; Chern et al., 2023).

Despite recent advances, existing methods often rely on feature concatenation or late-stage aggregation, and they struggle to model the complementary and conflicting relationships.

3 Method

As illustrated in Fig. 2, given a claim, we generate three analytical perspectives focusing on statement plausibility, retrieved-source credibility, and evidence consistency (Vlachos and Riedel, 2014) (see Section 3.1). The resulting multi-perspective analyses are then fed into an interaction module, which enables cross-perspective reasoning and aggregates them to reach a consensus for the final verification decision (see Section 3.2).

3.1 Multi-Perspective Analysis

In real-world fact-checking, human fact-checkers typically follow a general check-worthiness assessment pipeline (Soprano et al., 2024), that is, determining whether a claim warrants verification, evidence retrieval, truthfulness assessment, deliberation to reach a consensus, and finally assignment of a judgment.

Our framework is inspired by this pipeline. As shown in Fig. 3, given a claim, our framework first prompts an LLM to perform a statement plausibility analysis. Then it automatically retrieves relevant evidence via web search and employs LLMs to conduct a truthfulness assessment from retrieved-source credibility and evidence consistency.

Statement Plausibility Analysis. This perspective evaluates the intrinsic plausibility of a claim based on its internal linguistic and logical properties. Specifically, it examines whether the claim exhibits abnormal patterns, such as logical inconsistency or implausible causal relations. In this work, we utilize the intrinsic knowledge and reasoning abilities of LLMs to assess the logical coherence and plausibility of a claim in the real-world. This process produces an initial veracity judgment (i.e., real or fake) accompanied by a concise explanation, providing an intuitive reference that complements subsequent multi-perspective verification.

Retrieved-Source Credibility Analysis. The perceived credibility of a claim is often informed by the trustworthiness of the evidence sources that report or verify it. In professional fact-checking practice, information extracted from authoritative domains and established news outlets is generally regarded as a more reliable indicator for validation (Vlachos and Riedel, 2014). Following this principle, our framework incorporates the credibility of retrieved evidence sources as an indicative

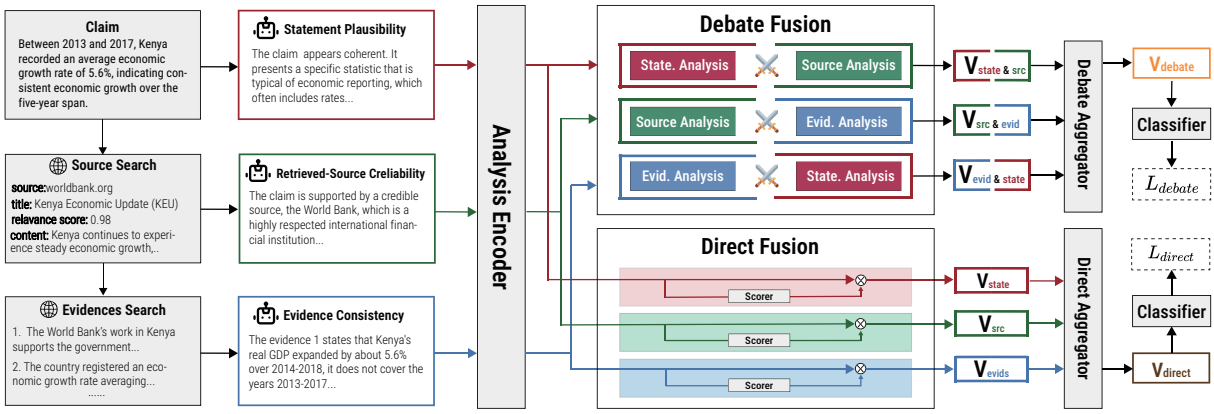

Figure 2: Architectural overview of the proposed framework. The pipeline initiates with automated evidence retrieval via web search based solely on the input claim. Subsequently, an LLM is employed to synthesize multi-perspective reasoning texts, capturing diverse evidential characteristics. The claim and its corresponding analyses are embedded using a shared encoder, forming three semantic sequences. These sequences are processed in parallel through two distinct components: a Debate Fusion module and a Direct Fusion module, generating feature representations V_{debate} and V_{direct}, respectively. Finally, these features are fed into their respective classifiers, where the individual losses are aggregated to form a joint objective function for model optimization.

signal, evaluating the prior credibility of each domain that provides the snippets.

For each claim, we retrieve the top-k most relevant results using the original claim as a query. From each retrieved result, we extract the *URL* as the source identifier, together with the *title* and the most relevant *content snippets*. By jointly conditioning on the claim and these related source elements, the LLM examines whether the reported content explicitly supports, contradicts, or ambiguously relates to the claim and assesses whether such evidence is trustworthy given the source's credibility. Based on this joint analysis, the LLM produces a source-informed veracity judgment accompanied by a concise explanation.

Evidence Consistency Analysis. This perspective provides a foundation for verification by identifying logical coherence or factual contradictions among multiple pieces of evidence. Since individual reports may be fragmented or biased, cross-referencing multiple sources helps mitigate hallucinations and increases the objectivity of the final judgment. We use web search tools to retrieve the top-n most relevant content snippets, which are organized into a structured evidence list. An LLM is then asked to perform a comparative analysis, reasoning about how each piece of evidence supports or contradicts the claim.

3.2 Debate Fusion Network

Building on the multi-perspective analyses generated by LLMs described in Section 3.1, we introduce a debate-inspired network to integrate these perspectives into a unified representation for the final judgment. The core idea of this network is to enable different perspectives to interact and "debate" with each other, encouraging complementary and integration.

Specifically, the network forms pairs from the three perspectives and, for each pair, dynamically selects the most suitable fusion strategy based on the consistency of their predictions regarding the retrieved evidence.

Representation. We employ the RoBERTa model as a shared analysis encoder. Given a claim and its three corresponding analytical texts (statement plausibility, retrieved-source credibility, and evidence consistency), we concatenate the claim with each analysis and feed the resulting sequence into the encoder to obtain their respective representations: $\mathbf{A}_{\text{state}}$, $\mathbf{A}_{\text{src}}$, and $\mathbf{A}_{\text{evid}}$. Each representation is defined as:

$$\mathbf{A}_a = \text{Encoder}([\text{claim}; \text{analysis}]), \mathbf{A}_a \in \mathbb{R}^{B \times L \times d_{\text{m}}} \quad (1)$$

Here, $\mathbf{A}_a$ denotes the contextualized sequence embeddings produced by the encoder, where $a \in \{\text{state}, \text{src}, \text{evid}\}$ represents the specific analytical aspect. The terms "claim" and "analysis" refer to the textual inputs corresponding to the claim and

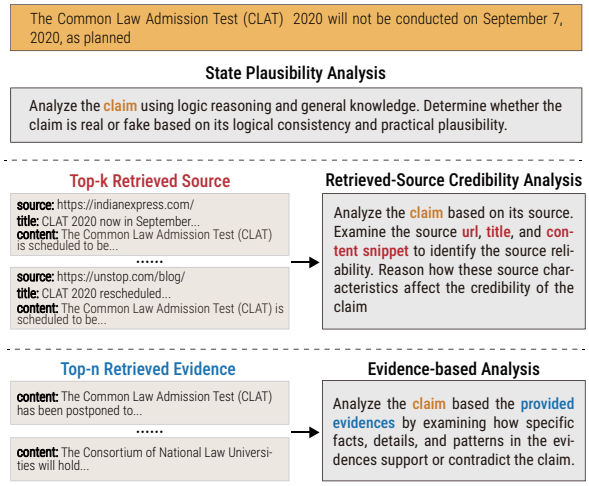

Figure 3: The analysis process of the originally retrieved evidence is as follows: (1) We use the claim itself to conduct statement plausibility analysis. (2) We extract the top-k pieces of evidence with the highest relevance scores and use their URLs, titles, and content snippets to construct a "micro news", prompting the LLM to examine this news according to its information source. (3) We extract content from the top-n pieces of evidence to form an evidence list, prompting the LLM to analyze the claim based on these pieces of evidence.

a specific analytical aspect, respectively. B is the batch size. L is the sequence length. d_{m} is the hidden dimensionality of the encoder.

3.2.1 Debate Fusion Module

The debate fusion module integrates multiple perspectives in an agreement-aware manner: it models interactions among perspectives and dynamically adjusts the aggregation process according to the level of agreement or conflict between them.

This module receives three analytical representations as input and constructs all pairwise combinations among them: $(\mathbf{A}_{\text{state}}, \mathbf{A}_{\text{src}})$, $(\mathbf{A}_{\text{state}}, \mathbf{A}_{\text{evid}})$, and $(\mathbf{A}_{\text{src}}, \mathbf{A}_{\text{evid}})$. For each pair of perspectives $(\mathbf{A}_a, \mathbf{A}_b)$, the module first applies average pooling along the sequence dimension to obtain condensed feature vectors:

$$\begin{aligned} \mathbf{h}_a &= \text{AvgPool}(\mathbf{A}_a), \quad \mathbf{h}_a \in \mathbb{R}^{B \times d_{\text{model}}} \\ \mathbf{h}_b &= \text{AvgPool}(\mathbf{A}_b), \quad \mathbf{h}_b \in \mathbb{R}^{B \times d_{\text{model}}} \end{aligned} \tag{2}$$

Then, for each pair, a shared multilayer perceptron (MLP) predicts the probability of the current claim being real or fake for both perspectives in the pair, denoted as p_a and p_b.

$$\begin{aligned} p_a &= \sigma(\text{MLP}(\mathbf{h}_a)), p_a \in [0, 1] \\ p_b &= \sigma(\text{MLP}(\mathbf{h}_b)), p_b \in [0, 1] \end{aligned} \tag{3}$$

Here, $\sigma(\cdot)$ is the activation function of the sigmoid. Each probability p_i represents the model's confidence in the truthfulness of the claim, as judged from the perspective i. A value of p_i closer to 0 means that the perspective i considers the claim to be more likely to be fake, while a value closer to 1 indicates greater confidence that the claim is real.

Analyses generated from different perspectives may yield consistent or conflicting predictions for the same claim. To quantify the extent of alignment between their judgments, we use L1 distance to calculate the consistency score based on their estimated probabilities p_a and p_b:

$$\text{score}_{(a,b)} = 1 - |p_a - p_b| \tag{4}$$

This metric intuitively measures the degree of agreement between two perspectives: the score increases as their predictions become more consistent for a given claim and decreases as their judgments diverge.

The consistency score serves as an adaptive signal for the subsequent fusion process, enabling the model to dynamically adjust how different perspectives are combined based on their pairwise agreement. We design two fusion strategies that operate according to the consistency level between perspective pairs.

Consensus Fusion Strategy. When two perspectives exhibit a high level of agreement, it implies an evaluative alignment regarding the claim's veracity. In this case, the two representations are integrated by element-wise addition:

$$\mathbf{v}_{\text{cons}} = \text{MaxPool}(\mathbf{A}_a) + \text{MaxPool}(\mathbf{A}_b) \tag{5}$$

Hence, their additive combination naturally reinforces the concordant evidence while preserving consistent information.

Divergence Fusion Strategy. The network utilizes a divergence strategy to handle the situation in which two perspectives are in disagreement. It uses a bi-directional cross-attention mechanism to reason over conflicting signals and extract informative context for subsequent decision making. First, we employ bi-directional cross attention to capture fine-grained interactions between the two texts, ensuring that each text can attend to important tokens in the other.

$$\begin{aligned} \mathbf{C}_{a \to b} &= \text{CrossAttn}(\mathbf{A}_a, \mathbf{A}_b, \mathbf{A}_b) \\ \mathbf{C}_{b \to a} &= \text{CrossAttn}(\mathbf{A}_b, \mathbf{A}_a, \mathbf{A}_a) \end{aligned} \tag{6}$$

After obtaining token-level cross representations, we apply max pooling along the sequence dimension to extract the most salient features and integrate the two representations by element-wise addition.

$$\mathbf{v}_{\text{div}} = \text{MaxPool}(\mathbf{C}_{a\to b}) + \text{MaxPool}(\mathbf{C}_{b\to a}) \tag{7}$$

The consistency score computed in the previous step is used as a differentiable soft gating weight to adaptively combine the outputs of the two fusion strategies. Specifically, for each evidence pair, we perform a weighted interpolation between the consensus and divergence strategies to obtain a fused debate vector, denoted $\mathbf{v}_{(a,b)}$:

$$\mathbf{v}_{(a,b)} = score_{(a,b)} \cdot \mathbf{v}_{\text{cons}} + (1 - score_{(a,b)}) \cdot \mathbf{v}_{\text{div}} \tag{8}$$

Finally, all three pairwise "debated" fusion results are aggregated using an attention mechanism to produce a final representation.

$$\begin{aligned} \mathbf{v}_{\text{debate}} &= \sum_{i\in\mathcal{P}} f_{\text{score}}(\mathbf{v}_i)\, \mathbf{v}_i, \\ \mathcal{P} &= \{(state, src), (state, evid), (src, evid)\}, \\ \hat{y}_{\text{debate}} &= f_{\text{MLP}}(\mathbf{v}_{\text{debate}}) \end{aligned} \tag{9}$$

This design enables the network to learn end-to-end how to adaptively transition from simple aggregation to complex interaction modeling, depending on the consistency among evidence.

3.2.2 Direct Fusion Module

In addition to the debate fusion module, we also introduce a direct fusion module that performs independent reasoning over each analysis. By processing each analysis separately, the model can extract salient features from individual analyses without being influenced by potentially noisy interactions, which can provide a complementary perspective to the debate network.

Each evidence's feature is pooled and scored via a learned linear layer. The scores are then normalized (e.g., via softmax) to produce weights:

$$\begin{aligned} \mathbf{c}_i &= \text{MaxPool}(\mathbf{C}_i), \\ \mathbf{s}_i &= f_{\text{score}}(\text{AvgPool}(\mathbf{C}_i)), i \in \{\text{state}, \text{src}, \text{evid}\} \end{aligned} \tag{10}$$

The final aggregated feature vector is obtained by weighting each evidence's pooled feature and summing:

$$\begin{aligned} \mathbf{v}_{\text{direct}} &= \sum_i s_i\, \mathbf{c}_i, \\ \hat{y}_{\text{direct}} &= f_{\text{MLP}}(\mathbf{v}_{\text{direct}}). \end{aligned} \tag{11}$$

The final loss function is computed using the cross-entropy function, measuring the differences between the true label of claim Y and the two predictions of the model $\hat{y}_{\text{debate}}$ and $\hat{y}_{\text{direct}}$:

$$\begin{aligned} L_{debate} &= \text{BCE}(Y, \hat{y}_{\text{debate}}), \\ L_{direct} &= \text{BCE}(Y, \hat{y}_{\text{direct}}), \\ L &= \gamma \cdot L_{debate} + (1-\gamma) \cdot L_{direct} \end{aligned} \tag{12}$$

Where γ is hyperparameter used to balance the two loss functions, and BCE is the binary cross-entropy loss function.

4 Experiments

4.1 Experimental Setup

Datasets. As shown in Table 2, we conducted experiments on three benchmark datasets: MMCoVaR (Chen et al., 2021), AVeriTeC (Schlichtkrull et al., 2023), and LIAR (Wang, 2017). For all datasets, we use only the claims and their corresponding true labels (restricted to real/fake values), without using any other dataset-specific information. MMCoVaR is released under the CC BY 4.0 license . The licenses for LIAR and Averitec were not explicitly stated on their official distribution pages. This approach enables us to evaluate our method in a pure claim verification setting, ensuring that the model's ability to judge claim veracity is based solely on the claim text.

Baselines. We compare three groups of methods: The **RoBERTa** baseline in the claim text without incorporating any external information (Liu et al., 2019). The **LLM** baseline employs **GPT-4o-mini** (OpenAI team, 2024) to reason about each claim from three different perspectives:

- **Statement Plausibility** assessing whether a claim is reasonable and logically coherent;
- **Retrieved-Source Credibility** incorporating retrieved source information to assess the credibility of the claim's origin;
- **Evidence Consistency** using the reasoning ability of LLMs to analyze multiple pieces of retrieved evidence.

LLM-Fusion uses an LLM to replace the debate fusion network, directly aggregating the three perspectives to produce the final verification verdict. **Bad Actor Good Advisor** (Hu et al., 2024) employs LLM to generate analyses from two perspectives and designed a network to classify the

Methods	MMCoVaR				AVeriTeC				LIAR			
	F1	Recall	Precision	Acc.	F1	Recall	Precision	Acc.	F1	Recall	Precision	Acc.
RoBERTa	0.851	0.850	0.853	0.862	0.735	0.737	0.736	0.766	0.642	0.648	0.650	0.643
Statement Plausibility	0.450	0.528	0.729	0.649	0.707	0.699	0.720	0.753	0.636	0.641	0.641	0.636
Retrieved-Source Credibility	0.746	0.734	0.776	0.778	0.729	0.753	0.726	0.744	0.643	0.658	0.667	0.645
Evidence Consistency	0.598	0.603	0.664	0.680	0.743	0.748	0.739	0.769	0.669	0.671	0.670	0.669
LLM-Fusion	0.738	0.737	0.738	0.756	0.694	0.679	0.770	0.769	0.646	0.649	0.668	0.663
Robust FND	0.833	0.829	0.840	0.847	0.717	0.715	0.720	0.757	0.527	0.566	0.525	0.572
Bad Actor Good Advisor	0.881	0.874	0.893	0.892	0.762	0.779	0.758	0.780	0.678	0.680	0.683	0.682
Covid-FND	0.879	0.875	0.886	0.889	0.795	0.805	0.789	0.814	0.697	0.700	0.701	0.698
Debate-to-Consensus (ours)	**0.902**	**0.897**	**0.907**	**0.909**	**0.818**	**0.830**	**0.811**	**0.834**	**0.736**	**0.736**	**0.736**	**0.737**

Table 1: Comparison of model performance on **MMCoVaR**, **AVeriTeC**, and **LIAR** datasets. The best results are highlighted in **Bold**, and the second-best results are Underline.

Dataset	Num. Claims
MMCoVaR (Chen et al., 2021)	2,593
AVeriTeC (Schlichtkrull et al., 2023)	2,589
LIAR (Wang, 2017)	3,671
Total	**8,853**

Table 2: Statistics of datasets used for our experiments.

claim according to the analyses. **CovidFND** (Irnawan et al., 2025), which also adopts the LLM + SLM framework, but additionally incorporates analyses of locally retrieved documents. **RobustFND** (Wu et al., 2024), which uses LLMs to generate semantic-flipped and paraphrased claims, thereby enriching the semantic diversity of the claims.

4.2 Training Settings

Following previous work (Xu et al., 2022), we employ 5-fold cross-validation to train and evaluate our models. Each dataset is partitioned into five folds, with four used for training and the remaining fold for testing. We train the models using the AdamW optimizer with an initial learning rate of 5e-5 for 50 epochs, applying early stopping to prevent overfitting. The batch size is set to 64 and the maximum input sequence length is limited to 500 tokens. A fixed random seed is used to ensure reproducibility. In our implementation, we adopt **GPT-4o-mini** as the LLM (OpenAI team, 2024) which sets temperature=0. We employ **Tavily** as the web search tool to retrieve relevant evidence (Tavily Team), which sets k=1 and n=5. We also activate its website blocking function to filter out pages that could cause data leakage (e.g., fact-checking interfaces or dataset pages) when searching with original claims (Tavily Team). All experiments were conducted on a single NVIDIA RTX A6000 GPU with 48 GB of memory. Evaluation metrics include F1 score, recall, precision, and accuracy.

4.3 Performance

Table 1 presents the performance of the proposed Debate-to-Consensus model compared to other baselines. The results lead to the following observations:

(1) Our method outperforms the RoBERTa baseline, achieving improvements of 5% in macro-F1, recall, and precision in MMCoVaR and 9% in AVeriTeC and LIAR. These results demonstrate the clear advantage of incorporating retrieved-source credibility and evidence consistency rather than relying solely on the claim text for verification.

4.4 Ablation Study

(2) Our method outperforms all LLM baselines, achieving an average improvement of 14% in all metrics compared to the best-performing LLM baseline in MMCoVaR , 8% in AVeriTeC, and 6% in LIAR. These results demonstrate that our method effectively integrates multiple perspectives to achieve more accurate judgments than using only an LLM, highlighting the advantage of using the LLM as an analyzer to enhance semantic representation while employing a dedicated network for classification.

(3) Our method consistently outperforms all the compared approaches in all four metrics, achieving gains of more than 2% on the MMCoVaR and AVeriTeC datasets, and more than 3% on the LIAR dataset. This demonstrates both the effectiveness and the robustness of our approach. Compared to approaches that rely solely on LLM-based analysis, our method incorporates explicit evidence retrieval to mitigate hallucinations. Moreover, rather than modeling each perspective in isolation, we

Methods	MMCoVaR				AVeriTeC				LIAR			
	F1	Recall	Precision	Acc.	F1	Recall	Precision	Acc.	F1	Recall	Precision	Acc.
Debate-to-Consensus (ours)	**0.902**	**0.897**	**0.907**	**0.909**	**0.818**	**0.830**	**0.811**	**0.834**	**0.736**	**0.736**	**0.736**	**0.737**
w/o Direct Fusion Module	0.836	0.845	0.813	0.844	0.808	0.816	0.803	0.827	0.732	0.732	0.733	0.734
w/o Debate Fusion Module	0.82	0.857	0.878	0.877	0.808	0.816	0.803	0.826	0.720	0.721	0.728	0.725
w/o Statement Plausibility	0.896	0.890	0.901	0.901	0.773	0.794	0.767	0.788	0.722	0.722	0.724	0.723
w/o Retrieved-Source Credibility	0.868	0.862	0.879	0.875	0.756	0.774	0.750	0.773	0.687	0.687	0.687	0.688
w/o Evidence Consistency	0.895	0.891	0.902	0.900	0.797	0.818	0.789	0.812	0.701	0.701	0.701	0.701

Table 3: Ablation results on **MMCoVaR**, **AVeriTeC**, and **LIAR**. The best results are shown in **Bold**, while the best result in each ablation group is Underline.

employ a debate fusion network that enhances perspective interaction and balances complementary and conflicting signals across perspectives. This design enables our approach to achieve the best overall performance.

We conducted ablation studies from two points of view to examine the effectiveness of different components in our framework. First, we ablate the fusion modules to analyze the contribution of each aggregation mechanism, including removing the direct fusion or the debate fusion module.

- w/o Direct Fusion Module: This variant removes the direct fusion module and relies solely on the output of the debate fusion module for prediction.
- w/o Debate Fusion Module: This variant removes the debate fusion module and relies solely on the output of the direct fusion module for prediction.

Second, we ablate each perspective to evaluate its individual impact on claim verification.

- w/o Statement Plausibility: This variant removes the statement plausibility analysis.
- w/o Retrieved-Source Credibility: This variant excludes retrieved-source credibility analysis.
- w/o Evidence Consistency: This variant removes evidence consistency analysis.

The results is shown in Table. 3.

The ablation result in fusion modules shows that the model using only the debate fusion module outperforms the variant using only the direct fusion module, indicating that the debate process allows the model to capture richer information than direct reasoning alone. The model achieves performance comparable to the complete architecture on LIAR but its performance drops noticeably on the other two datasets. In contrast, the complete framework consistently achieves the best results across most evaluation metrics on all three datasets, indicating that the two modules are complementary and jointly contribute to improved performance and greater stability.

The ablation result in perspectives shows that removing any single perspective consistently degrades model performance, confirming the complementary nature of the proposed multi-perspective design. Among the three perspectives, eliminating retrieved-source credibility leads to the largest performance drop across all three datasets, highlighting the critical role of source-level credibility. Moreover, statement plausibility has a greater impact than evidence consistency on AVeriTeC, whereas the opposite trend is observed on LIAR. This observation suggests that claims with different characteristics may benefit from different analytical perspectives, integrating complementary perspectives enables our model to adapt to a wider range of claims and achieve more robust verification performance.

5 Conclusion

In this paper, we propose Debate-to-Consensus, a claim verification framework that uses web search to automatically collect relevant evidence for a given claim and supports multi-perspective debate reasoning to assess claim reliability. Experimental results on three benchmark datasets demonstrate that our approach consistently outperforms existing baselines, validating the effectiveness of multi-perspective analysis and the importance of modeling inter-perspective interactions in achieving more reliable and explainable claim verification.

6 Limitation

Despite its effectiveness, our framework currently only outputs a binary veracity judgment (true or false); the final judgment is not yet presented in

an interactive manner. An important future direction is to extend the framework into an interactive fact-checking tool that automatically retrieves evidence, outputs veracity judgments, and visualizes agreements and conflicts among perspectives and evidence sources, thereby making the decision process more transparent and actionable for human users.

Although the proposed perspectives based on intuition, source and content capture the core human verification strategies, they do not include all perspectives necessary for complex claim verification. Certain claims may require specialized reasoning that is outside the current set of perspectives. However, naively expanding the number of analytical perspectives does not inherently guarantee performance gain.Based on empirical evaluation and trade-off analysis, we select these three complementary perspectives to achieve an effective balance between coverage and robustness. In future work, we aim to explore additional perspectives to further enhance claim verification, investigating how specialized reasoning and new angles can improve the robustness and accuracy of the verification process.

Although we use a web search for automatic evidence collection, Debate-to-Consensus is still limited by the quality and coverage of the retrieved evidence. Missing, noisy, or biased results during the retrieval process can reduce the reliability of verification, particularly for emerging or niche claims. To address this concern, we retain an intuition-based analysis perspective that leverages the intrinsic knowledge of LLMs as a complementary signal when the retrieved evidence is unreliable or insufficient. In future work, we plan to incorporate fine-grained curated knowledge bases and explicit evidence validation mechanisms to further improve retrieval quality and robustness.

References

Mingxuan Chen, Xinqiao Chu, and K. P. Subbalakshmi. 2021. MMCoVaR: multimodal COVID-19 vaccine focused data repository for fake news detection and a baseline architecture for classification. In *Proc. Int. Conf. Advances in Social Networks Analysis and Mining, Virtual Event (ASONAM)*, pages 31–38. ACM.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Tsun-Hin Cheung and Kin-Man Lam. 2023. FactLLaMA: Optimizing instruction-following language models with external knowledge for automated fact-checking. In *Proc. IEEE Asia Pacific Signal and Information Processing Association Annual Summit and Conf.*, pages 846–853.

Maryanne Garry, Way Ming Chan, Jeffrey Foster, and Linda A. Henkel. 2024. Large language models (llms) and the institutionalization of misinformation. *Trends in Cognitive Sciences*, 28(12):1078–1088.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proc. AAAI Conf. Artificial Intelligence and Conf. on Innovative Applications of Artificial Intelligence and Symp. on Educational Advances in Artificial Intelligence*, pages 22105–22113.

Bassamtiano Renaufalgi Irnawan, Sheng Xu, Noriko Tomuro, Fumiyo Fukumoto, and Yoshimi Suzuki. 2025. Claim veracity assessment for explainable fake news detection. In *Proc. Int. Conf. on Computational Linguistics (COLING)*, pages 4011–4029.

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

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)*, pages 15556–15583.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with LLM enhanced semantics mining. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pages 508–521.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proc. Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1732–1742.

OpenAI team. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proc. Conf. on Empirical Methods in Natural Language Processing*, pages 22–32.

Simeng Qin and Mingli Zhang. 2024. Boosting generalization of fine-tuning BERT for fake news detection. *Information Processing & Management*, 61(4):103745.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive biases in fact-checking and their countermeasures: A review. *Information Processing Management*, 61(3):103672.

Xin Tan, Bowei Zou, and Ai Ti Aw. 2025. Improving explainable fact-checking with claim-evidence correlations. In *Proc. Int. Conf. on Computational Linguistics (COLING)*, pages 1600–1612.

Tavily Team. Tavily: Real-time web search api for ai agents. https://tavily.com/.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proc. ACL Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: generating reactions and explanations for llm-based misinformation detection. In *Findings of the Association for Computational Linguistics (ACL)*, pages 2637–2667.

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)*, pages 6288–6304.

William Yang Wang. 2017. "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 422–426.

Wenjie Wei, Yanyue Zhang, Jinyan Li, Panfei Liu, and Deyu Zhou. 2025. Cross-domain fake news detection based on dual-granularity adversarial training. In *Proc. Int. Conf. on Computational Linguistics (COLING)*, pages 9407–9417.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics (ACL/IJCNLP)*, pages 2560–2569.

Yike Wu, Yang Xiao, Mengting Hu, Mengying Liu, Pengcheng Wang, and Mingming Liu. 2024. Towards robust evidence-aware fake news detection via improving semantic perception. In *Proc. Joint Int. Conf. on Computational Linguistics, Language Resources and Evaluation (LREC/COLING)*, pages 16607–16618.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proc. ACM Web Conf.*, pages 2501–2510.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10331–10343.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yuming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics (ACL)*, pages 9274–9281.

Xiaofan Zheng, Minnan Luo, and Xinghao Wang. 2025. Unveiling fake news with adversarial arguments generated by multimodal large language models. In *Proc. Int. Conf. on Computational Linguistics*, pages 7862–7869.