Debating Truth: Debate-driven Claim Verification with Multiple Large Language Model Agents

Haorui He[1,2], **Yupeng Li**[1,*], **Dacheng Wen**[1,2], **Reynold Cheng**[2], **Francis C. M. Lau**[2]
[1]Department of Interactive Media, Hong Kong Baptist University
[2]Department of Computer Science, The University of Hong Kong

Abstract

Claim verification is critical for enhancing digital literacy. However, the state-of-the-art single-LLM methods struggle with complex claim verification that involves multi-faceted evidences. Inspired by real-world fact-checking practices, we propose *DebateCV*, the first claim verification framework that adopts a debate-driven methodology using multiple LLM agents. In our framework, two *Debaters* take opposing stances on a claim and engage in multi-round argumentation, while a *Moderator* evaluates the arguments and renders a verdict with justifications. To further improve the performance of the Moderator, we introduce a novel post-training strategy that leverages synthetic debate data generated by the zero-shot DebateCV, effectively addressing the scarcity of real-world debate-driven claim verification data. Experimental results show that our method outperforms existing claim verification methods under varying levels of evidence quality. Our code and dataset are publicly available at `https://anonymous.4open.science/r/DebateCV-6781`.

1 Introduction

In the modern digital landscape where misinformation disseminates widely and rapidly online, (automated) claim verification has become a cornerstone technique for assessing the veracity of information, which mitigates the influence of misinformation and enhances digital literacy (Vlachos and Riedel, 2014). Existing state-of-the-art (SOTA) claim verification methods typically leverage a large language model (LLM) to evaluate the veracity of a claim by analyzing its consistency or contradiction with any relevant evidence collected from external sources, e.g., the web, via retrieval-augmented generation (RAG) (Schlichtkrull et al., 2024a). In this framework, the target claim and the collected evidences are integrated into a structured prompt that directs a single LLM to generate a predicted verdict with corresponding justifications, which elucidate the rationale behind its veracity assessment.

However, individual LLMs are limited when verifying complex, multi-faceted claims. Take *HerO* (Yoon et al., 2024), the SOTA that leverages the above framework with a single LLM and RAG, as an example.[1] It incorrectly refutes the claim "`52% of Nigeria's current population lives in urban areas`," by relying on an outdated evidence indicating a 36% urban population in 2008. Notably, a more recent evidence explicitly supporting a 52% urbanization rate was also successfully retrieved, yet the single-LLM solution failed to reconcile contradictions across the collected evidences.

In real-world practices, fact-checking organizations commonly adopt a **debate-driven claim verification** methodology to address errors in individual assessments. For example, Graves (2013) describes how PolitiFact, a Pulitzer Prize-winning fact-checking organization, employs "star chamber" sessions, i.e., debates among fact-checkers, for claim verification. In such sessions, a panel of fact-checkers critically evaluates each other's veracity assessments to identify potential flaws in the evidence analysis, e.g., the oversight of the latest urbanization statistics in the above example. This adversarial process compels fact-checkers to address identified weaknesses, re-analyze evidences, and strengthen assessments through iterative refinement.[2] Such a debate-driven claim verification methodology enhances the depth and rigor of evidence analysis, fostering holistic claim evaluation.

Inspired by the above, we propose **DebateCV**, the first claim verification framework that adopts such a debate-driven methodology using multiple LLM agents. The framework involves two

[*] Correspondence: ivanypli@gmail.com

[1] This error traces back to Claim 218 in Yoon et al. (2024)'s official results. Full details are available in our repository.

[2] For more specific examples of these real-world claim verification debates, see pages 204–208 of (Graves, 2013).

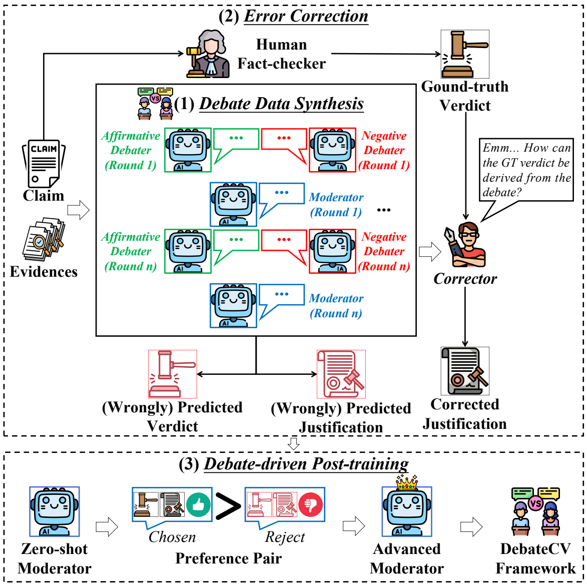

(2) Error Correction
Human Fact-checker
Gound-truth Verdict
(1) Debate Data Synthesis
Claim
Evidences
Affirmative Debater (Round 1)
Negative Debater (Round 1)
Moderator (Round 1)
Affirmative Debater (Round n)
Negative Debater (Round n)
Moderator (Round n)
Emm... How can the GT verdict be derived from the debate?
Corrector
(Wrongly) Predicted Verdict
(Wrongly) Predicted Justification
Corrected Justification
(3) Debate-driven Post-training
Zero-shot Moderator
Chosen
Reject
Preference Pair
Advanced Moderator
DebateCV Framework

Figure 1: An overview of DebateCV with synthetic debate-driven claim verification data for post-training.

agents, referred to as the *Debaters*, taking opposing stances: one supporting and the other refuting the claim. They are instructed to challenge each other's positions and defend their own using the collected evidences across several rounds. In each round, a third agent, the *Moderator*, evaluates the arguments put forth by the Debaters, summarizes the current round of debate, and determines whether further debate is required. If not, the Moderator issues a final verdict and generates justifications. By structuring the claim verification as an adversarial "star chamber" among LLM agents, our framework analyzes evidence from opposing perspectives, ultimately reaching a well-supported verdict.

The performance of DebateCV then hinges on the ability of the Moderator to evaluate arguments during multi-agent debates and determine which side has stronger, evidence-grounded arguments. Post-training, shown to improve the performance of LLMs across various tasks (Tie et al., 2025), offers a promising approach to enhance this capability. Applying post-training to debate-driven claim verification requires annotated datasets containing debate recordings, i.e., the (multi-round) argument logs that lead to the final verdicts for each claim. Unfortunately, such real-world data is not publicly available and collecting it via human annotation is time-consuming and costly.

To address this data scarcity, we propose a novel post-training strategy for debate-driven claim verification that leverages synthetic data. Specifically, we utilize the zero-shot DebateCV to simulate the process of debating claims with human-labeled verdicts and evidences, thereby synthesizing the debate recordings by the Debaters and pairing them with verdicts and justifications from the Moderator. Since the zero-shot Moderator may generate (erroneous) justifications that lead to incorrect verdicts, we introduce an additional LLM agent, the *Corrector*, which detects discrepancies between the predicted justifications and the ground-truth verdicts, and revises them accordingly to align with the ground-truth. By leveraging such synthetic data, we design a tailored debate-driven supervised fine-tuning (D-SFT) and direct preference optimization (D-DPO) to post-train the Moderator (see Sec. 4 for details). Our experiments show that this strategy significantly improves the Moderator for producing accurate verdicts in our debate-driven framework.

Fig. 1 illustrates an overview of our methods. Our contributions are summarized as follows.

(1) We propose DebateCV, the first claim verification framework that adopts a debate-driven methodology using multiple LLM agents.

(2) We propose a post-training strategy for debate-driven claim verification leveraging synthetic debate data.

(3) Experimental results show that our method significantly outperforms existing methods in both golden and retrieved evidence settings, with accuracy improving by 3.0% and 1.8%, respectively.

2 Related Works

Claim Verification. The state-of-the-art claim verification methods (Rothermel et al., 2024; Yoon et al., 2024; Ullrich et al., 2024; Wang et al., 2024) typically rely on the RAG framework, prompting single LLMs with claims and evidences to predict verdicts from one of the following categories: "Supported," "Refuted," "Not Enough Evidence," or "Conflicting Evidence/Cherry-picking," along with generating justifications.[3] However, recall our example in Sec. 1, a single LLM struggles to consider nuanced details that are necessary to verify complex multi-faceted claims. This is likely due to the mismatch between LLMs' pretraining objective of learning linguistic patterns and autoregressively predicting the most probable next token and the demands of claim verification require nuanced analysis of evidences and claims. As discussed earlier, some real-world fact-checking orga-

[3] See Appendix A for the verdict category definitions.

nizations leverage critical debate over fine-grained details in the collected evidence to identify and correct potential errors in individual fact-checkers' assessments (Graves, 2013). Our proposed DebateCV framework, inspired by such real-world fact-checking practices, is the first automatic claim verification framework that leverages such a debate-driven methodology. Close to our work, Wang et al. (2024) highlight incorporating information from opposing perspectives to address misinformation. However, they primarily focus on the *evidence retrieval*, where evidence is retrieved separately for the two opposing sides (support or refute). In contrast, evidence retrieval is beyond the scope of this work. We focus on integrating opposing viewpoints through LLM-stimulated debate to *claim verification* with varying evidence quality.

Additionally, existing RAG-based methods, such as (Yoon et al., 2024; Ullrich et al., 2024), rely on standard supervised fine-tuning (SFT) to post-train LLMs to predict verdicts annotated by human fact-checkers, using a prompt that include claims and their relevant evidences. Our work proposes a novel post-training strategy for debate-driven claim verification that incorporates synthetic debate recordings as contexts. This strategy leverages the advantages of debate dynamics, leading to improved claim verification performance.

Multi-agent Debate. Recent studies have explored how multi-agent interactions among LLMs can enable various applications, such as AI societies (Li et al., 2023a) and software development (Wu et al., 2024; Hong et al., 2024). Several works have proposed multi-agent debate frameworks for specific downstream tasks, including machine translation (Liang et al., 2024), hallucination mitigation (Du et al., 2024), weak-to-strong supervision (Khan et al., 2024), and stance detection (Lan et al., 2024). These studies demonstrate that structured debate mechanisms often outperform alternative forms of multi-agent interaction in complex tasks (Liang et al., 2024; Du et al., 2024). Close to our work, Kim et al. (2024) focused on leveraging multi-agent debate to enhance the faithfulness of justifications, which represents only a subsequent step after claim verification. Jeptoo and Sun (2024) utilized multi-agent debate for fake news detection, a distinct task aimed at verifying the veracity of news based on linguistic or multi-modal patterns, without retrieving and analyzing external evidences. However, such patterns often capture only the characteristics of fake news within specific platforms (Li et al., 2024b) or constrained temporal windows (Yang et al., 2024), limiting the real-world applicability of these methods. In contrast, our claim verification task emulates a more realistic scenario akin to real-world fact-checking, which emphasizes the identification of contradictions or consistencies between collected evidences and the claim (Guo et al., 2022). To our best, we are the first to address claim verification via a multi-agent debate framework.

In addition, the above works focus on applying zero-shot LLMs in multi-agent debates, while only three prior studies leverage multi-agent debate to train LLMs. For instance, Subramaniam et al. (2024) use multi-agent debate to refine the training corpus of LLMs. Li et al. (2024a) employs multi-agent debates to generate a corpus of convincing arguments to fine-tune the persuasiveness of LLMs. Estornell et al. (2024) employ a debate game between two teams of LLM agents to enhance their team-working abilities. Different from these works, our work proposes a new paradigm to agents in multi-agent debates for a downstream task. We address that, though designed for our claim verification task, such a strategy, with minimal modifications, can be adapted to other contexts where multi-agent debate is performed.

3 The DebateCV framework

This section introduces DebateCV, which utilizes LLM agents to perform debate-driven claim verification through two key stages: (1) Pre-Debate Configuration and (2) In-Debate Process.

Pre-debate Configuration. Similar to debates conducted by humans, a debate needs to be well configured through role assignment. The debate in our proposed framework is performed by three LLM agents: two Debaters (*Affirmative* and *Negative*) and one Moderator. These roles are assigned through a meta prompt, which instructs actions of the agents throughout the multi-round debate. Their assigned responsibilities are outlined as follows: (1) The Debaters are tasked with verifying the veracity of a specific textual claim c. The Affirmative Debater supports the veracity of the claim while the Negative Debater refutes it. Both Debaters are provided with the same set of evidence $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$, which they can use to support their arguments in the debate. Each Debater is responsible for defending their positions with presented evidences, and critically analyzing and challenging the opposing argument based on the provided evidences. (2) The Moderator oversees the debate process, evaluating the strength of the evidence and arguments presented by the Debaters. Ultimately, the Moderator predicts the veracity $\hat{y}$ of claim c and generates a corresponding justification $\hat{j}$ based on the debate.

In-Debate Process. After the configuration, the agents start debating a presented claim in rounds. In the first round, the Affirmative Debater initiates the debate by presenting arguments based on the evidences to support claim c. In rebuttal, the Negative Debater presents counter-arguments, which can involve highlighting counter evidences, questioning the sufficiency of the Affirmative's evidences, and offering alternative explanations for why claim c should be refuted. In subsequent rounds, the Affirmative and Negative agents continue to defend their respective positions and challenge each other. At the end of each round, the Moderator summarizes the key arguments from both sides, identifies the primary points of contention, and evaluates whether the debate should proceed. If the arguments and points of contention remain unchanged, the debate is seen as converged and the Moderator terminates the debate early. Otherwise, the Moderator opts to let the debate continue. Once the debate ends, the Moderator outputs a final predicted verdict $\hat{y}$ based on the overall strength of the evidence-based arguments made in the debate, and generates a justification $\hat{j}$ for the decision.

The specific prompts used in the framework are available in the Appendix B.

4 Post-training the Moderator with Synthetic Debate Data

While pre-trained LLMs demonstrate strong zero-shot capabilities, adapting them through task-specific post-training can significantly improve their performance on downstream applications (Tie et al., 2025). In our preliminary experiments, we observe that while the Debaters generated well-reasoned arguments, the Moderator struggled with veracity assessments for complex claims requiring extended debate, and occasionally exhibited conformity bias (see Sec. 5.3). Therefore, we aim to improve its ability to evaluate debate dynamics and render accurate judgments through post-training. However, acquiring human-annotated data, which contains (multi-round) debate recordings conducted by fact-checkers that culminate in final verdicts for each claim, for such post-training is both time-consuming and costly. In response, this section describes our proposed post-training strategy based on synthetic debate data. Specifically, we synthesize debate recordings for claims in existing datasets that provide only human-annotated verdicts and evidence. As illustrated in Fig. 1, the method comprises three key steps: (1) Debate Data Synthesis, (2) Error Correction, and (3) Debate-driven Post-Training.

Debate Data Synthesis. Given the (zero-shot) DebateCV framework introduced in Sec. 3, we first use it to generate debate recordings and output verdicts with justifications. Specifically, given n claims $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ with their corresponding ground-truth verdicts $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ and human-annotated evidences $\mathcal{E} = \{\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n\}$, the DebateCV initiates debates $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ for verifying $\mathcal{C}$. The Moderator then provides the initial predicted veracity labels $\hat{\mathcal{Y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$, along with their justifications $\hat{\mathcal{J}} = \{\hat{j}_1, \hat{j}_2, \ldots, \hat{j}_n\}$.

Error Correction. The verdicts and justifications predicted by the zero-shot Moderator may contain inaccuracies. To address these errors, we introduce an additional LLM agent, Corrector, who generates revised justifications aligned with the ground-truth verdicts. Specifically, for each claim c_i paired with an wrongly predicted verdict $\hat{y}_i \neq y_i$, we provide the debate recording d_i and the ground-truth verdict y_i as input to the Corrector. We then instruct it to generate a corrected justification j'_i that explains how the ground-truth verdict y_i could be derived from the debate d_i. The specific prompt used by the Corrector is presented in Appendix B. This approach guides the Corrector to analyze the strengths and weaknesses of the arguments presented by both Debaters in d_i and attempts to articulate a coherent rationale supporting the ground-truth verdict. Since the evidence and the verdict are human-annotated, the corrected justification j'_i provides a factually grounded explanation of how the verdict follows from the debate contents.

We synthesize debates for the training split of AVeriTeC (Schlichtkrull et al., 2024b), the state-of-the-art dataset for claim verification, to create our synthetic debate dataset, **SynDeC**. Each sample in the dataset contains: Claim (c_i), Evidence Set ($\mathcal{E}_i$), Debate Recording (d_i), Ground-truth Verdict (y_i);

Predicted Verdict ($\hat{y}_i$); Predicted Justification ($\hat{j}_i$) and Corrected Justification (j'_i), if $\hat{y}_i \neq y_i$.

Debate-driven Post-training. Leveraging SynDeC, we propose a debate-driven post-training strategy that builds upon the standard LLM post-training paradigm, supervised fine-tuning (SFT) followed by reinforcement learning (RL) (Lin et al., 2024; Wang et al., 2023; Tie et al., 2025), while being specifically tailored to enhance the Moderator's performance in debate-driven claim verification.

First, we categorize the claims in SynDeC based on the accuracy of the predicted verdicts produced by the zero-shot Moderator: $\mathcal{C}_{\text{correct}} = \{c_i \mid \hat{y}_i = y_i\}$ and $\mathcal{C}_{\text{error}} = \{c_i \mid \hat{y}_i \neq y_i\}$. Then, we perform SFT on the Moderator using $\mathcal{C}_{\text{correct}}$ to enhance its claim verification performance. Specifically, treating the debate recording d_i as the context of a multi-turn dialogue, the Moderator is trained to generate responses that contain correct verdicts $\hat{y}_i = y_i$ and their justifications j'_i in the final round of the debate for each claim c_i. We refer to this process as *Debate-driven SFT (D-SFT)*, which results in an intermediate model:

$$\theta_{\text{SFT}} = \arg\min_{\theta} \mathcal{L}_{\text{SFT}}^{\text{correct}}(\theta),$$

where $\mathcal{L}_{\text{SFT}}$ is the SFT loss (Ouyang et al., 2022) and θ represents the zero-shot Moderator.

Next, during the RL phase, we apply the widely adopted Direct Preference Optimization (DPO) (Rafailov et al., 2024) to θ_{SFT} to correct the Moderator from its previous errors. Specifically, for each c_i in $\mathcal{C}_{\text{error}}$, we define the response containing the ground-truth verdict and corrected justification (y_i, j'_i) as the chosen response r^+, and the initial response with the incorrect verdict and justification $(\hat{y}_i, \hat{j}_i)$ as the rejected response r^-, forming a preference pair (r^+, r^-). DPO maximizes the probability of the chosen response r^+, while minimizing the probability of the rejected response r^-, thereby enhancing the claim verification performance of the final *Advanced Moderator* parameterized by

$$\theta_{\text{SFT w/ DPO}} = \arg\min_{\theta} \mathcal{L}_{\text{DPO}}(\theta_{\text{SFT}}),$$

where $\mathcal{L}_{\text{DPO}}$ is the DPO loss (Rafailov et al., 2024). We term this step as *Debate-driven DPO (D-DPO)*.

By combining both D-SFT, which strengthens accurate assessments, and D-DPO, which corrects erroneous assessments, our post-training progressively evolves the Moderator into an Advanced Moderator, enabling accurate debate-driven claim verification through debate evaluation.

5 Experiments

In this section, we present experiments evaluating our proposed methods (Sec. 5.2), and analyze how our method improves performance (Sec. 5.3).

5.1 Experimental Setups

Datasets and Evidence Conditions. We use AVeriTeC (Schlichtkrull et al., 2024b), the state-of-the-art real-world claim verification dataset. AVeriTeC consists of English textual claims collected from 50 fact-checking organizations and annotated by human fact-checkers, and addresses key limitations of previous datasets, including evidence leakage and insufficiency. It provides a knowledge base comprising evidences from approximately 1,000 online sources for each claim.

To systematically evaluate our methods under varying evidence conditions, we employ three types of evidences from the knowledge base: (1) *Golden Evidence*: We utilize the golden evidences employed by human fact-checkers. This condition represents the best performance achievable by any method; (2) *Retrieved Evidence*: To simulate a realistic end-to-end claim verification pipeline, we use an automatic evidence retrieval method that mimics open-web evidence retrieval, which may introduce noise (e.g., irrelevant evidence). Specifically, we adopt the best-performing retrieval strategy (Yoon et al., 2024) from the AVeriTeC challenge (Schlichtkrull et al., 2024a) (3) *No Evidence*: To evaluate whether DebateCV enhances evidence analysis, we introduce an ablation setup, in which no evidence is provided to the models. We fix the same set of evidence across all conditions to isolate its influence on the experiments.[4]

For training, we use the train split of AVeriTeC with with 3,068 claims: 2148 for D-SFT and 920 for D-DPO. For evaluation, we use its development split, which includes 500 claims.[5]

Baselines. We evaluate our method against both single-agent and multi-agent baselines, as well as ablation configurations.

Single-agent methods include: (1) *RAG-COT*: The state-of-the-art approach that prompts a single LLM to use provided evidences for verification. The method was employed by the top three solutions in the AVeriTeC Challenge (Rothermel

[4] Note that evidence retrieval is beyond our scope.

[5] The golden evidence and ground-truth labels for the test split of AVeriTeC are not publicly accessible. The evaluation data is unseen during training to prevent any potential bias.

et al., 2024; Yoon et al., 2024; Ullrich et al., 2024), albeit with variations in evidence retrieval strategies, prompting methods, and LLM backbones. In our experiments, we implement chain-of-thought (CoT) prompting strategy from Yoon et al. (2024), using GPT-4o (OpenAI, 2024) and Llama-3.1-8B-Instruct (Meta, 2023) (referred to as Llama-3.1) as the backbones. (2) *RAG-SFT*: An enhanced version of the RAG-COT baseline, where the LLM undergoes standard SFT for claim verification. We leverage the open-source checkpoint based on Llama-3.1 provided by Yoon et al. (2024).

Multi-agent methods include: (1) *Majority*: This baseline follows the approach of Lin et al. (2025), in which three independent RAG-COT-based agents verify each claim individually. The final verdict is determined via majority voting among the three agents, using the voting prompt from their official implementation. (2) *Cooperation*: This baseline adapts the CAMEL framework (Li et al., 2023a), which simulates collaborative interactions between an AI user and an AI assistant, to perform claim verification using the same prompting strategy as RAG-COT. Both GPT-4o and Llama-3.1 are used as backbones for all multi-agent baselines.

Ablational configurations include: (1) *DebateCV w/o D-SFT & D-DPO*: A variant of our framework without any post-training for the Moderator. For computational efficiency, we use GPT-4o-mini, as the Debaters, while the Moderator is based on GPT-4o and Llama-3.1. (2) *DebateCV w/o D-DPO*: A variant where the Moderator (Llama-3.1) undergoes only D-SFT, excluding the Corrector to construct preference pairs for D-DPO. (3) *DebateCV*: The full framework, where the Moderator (Llama-3.1) is post-trained using both D-SFT and D-DPO.

Other implementation details of the baselines, such as hyper-parameters and training strategies, are presented in Appendix C.

Metrics. We adopt both overall accuracy (Acc.) and the AVeriTeC score (AVer.) (Schlichtkrull et al., 2024b), which is accuracy when evidence is sufficient, for comprehensive evaluation. The AVeriTeC score first evaluates evidence quality using the METEOR score $f(\hat{\mathcal{E}}, \mathcal{E})$, where f denotes the scoring function that compares the (collected) evidences $\hat{\mathcal{E}}$ and the ground-truth evidences $\mathcal{E}$. A threshold of $f(\hat{\mathcal{E}}, \mathcal{E}) \geq 0.25$ is then applied to determine whether adequate evidences are collected. The AVeriTeC score is the accuracy of the predicted claim verdicts where the evidence score exceeds this threshold. Notably, AVeriTeC score is equivalent to accuracy in the golden evidence condition, and is constant zero in the no-evidence condition.

Following Chen et al. (2024), we use paired bootstrap tests to assess statistical significance based on at least five repeated trials. We use "†" to indicate statistical significance ($p \leq 0.05$) compared to the results of RAG-SFT.

5.2 Experimental Results

Main Results. Table 1 demonstrates that our proposed DebateCV consistently outperforms all baselines across various evidence conditions: (1) Under the golden evidence condition, DebateCV achieves a new state-of-the-art performance for claim verification, significantly improving accuracy from 80.4 (RAG-SFT) to 83.4. This result highlights the effectiveness of our debate-driven framework in enhancing the best performance achievable when high-quality evidences are available. (2) Under the more realistic retrieved evidence condition, DebateCV demonstrates superior robustness in end-to-end pipelines involving noisy, imperfectly retrieved evidence. It improves accuracy from 70.2 to 72.0 and increases the AVeriTeC score from 52.6 to 54.4 compared to RAG-SFT, reflecting improved resilience to suboptimal evidence quality.

Ablation Study. Our ablation analysis further reveals the contribution of each component of our framework: (1) **Impact of D-DPO**: DebateCV w/o D-DPO exhibits significant performance drops, particularly under the retrieved evidence condition. Accuracy decreases from 72.0 to 67.2, and the AVeriTeC score falls from 54.4 to 50.8. This underscores the importance of the Corrector in enabling preference learning via D-DPO, which rewards factually grounded justifications and penalizes initially erroneous predictions. (2) **Impact of D-SFT**: Despite being inferior to DebateCV, the w/o DPO variant still significantly outperforms the zero-shot DebateCV w/o D-SFT & D-DPO. This indicates that the D-SFT stage alone contributes meaningfully to overall performance improvements. (3) **Role of Evidence Analysis**: All DebateCV variants exhibit a notable performance drop under the no-evidence condition. For instance, in DebateCV w/o D-SFT & D-DPO, accuracy for GPT-4o drops from 57.4 (retrieved evidence) to 41.6 (no evidence). This indicates that a substantial portion of the performance gains in DebateCV stems from the analysis of evidence during multi-agent debates.

Methods	LLM	Golden		Retrieved		No-evidence	
		Acc.	AVer.	Acc.	AVer.	Acc.	AVer.
RAG-COT	GPT-4o	64.2	64.2	52.2	39.0	34.4	0.0
	LLama-3.1	61.4	61.4	45.6	36.8	25.8	0.0
RAG-SFT	LLama-3.1	80.4	80.4	70.2	52.6	57.6	0.0
Majority	GPT-4o	71.2	71.2	54.4	40.4	47.6	0.0
	LLama-3.1	62.0	62.0	40.2	30.0	29.2	0.0
Cooperation	GPT-4o	75.2	75.2	57.0	44.2	40.6	0.0
	LLama-3.1	62.0	62.0	45.4	34.4	35.8	0.0
DebateCV	LLama-3.1	**83.4**†	**83.4**†	**72.0**†	**54.4**†	**65.0**†	0.0
DebateCV w/o D-DPO	LLama-3.1	81.2†	81.2†	67.2	50.8	46.8	0.0
DebateCV w/o D-SFT & D-DPO	GPT-4o	75.8	75.8	57.4	42.8	41.6	0.0
	LLama-3.1	68.6	68.6	53.4	40.6	46.0	0.0

Table 1: Claim verification performance of different baselines under various evidence conditions.

Claim (ID; Label)	RAG-COT	Majority	DebateCV		
			Affirmative Debater	Negative Debater	Moderator
Amy Coney Barrett was confirmed as US Supreme Court Justice on Oct. 26, 2020. **(ID: 31; Supported)**	Most evidence supports Oct. 26, but some also say Oct. 27. This leads to Conflicting Evidence. ✗	Agent 3 noted Oct. 27 as the assumed duty date, and all agents confirm most evidence supports Oct. 26 as the confirmation date. ✓	Multiple sources state she was confirmed as a Supreme Court Justice on Oct. 26 via a 52–48 vote. Supported.	The term “confirmed” can be interpreted as the date she officially assumed duty, which is Oct. 27. Refuted.	The claim refers to the confirmation of her role, not the official swearing-in. The claim is supported. ✓
New Zealand’s Abortion Act does not mandate medical support for babies born alive after an abortion. **(ID: 99; Refuted)**	Multiple evidence confirms that an amendment to the Act proposing this duty was rejected. The claim is supported. ✗	While Agent 2 noted a provision requiring medical support, all agents confirmed the amendments were not passed with multiple evidences. ✗	Multiple evidence confirm there is no such legal duty; the rejection of the amendments is also a supporting evidence. Supported.	The Act clearly states that "... have a duty to provide medical care." The amendments were rejected as redundant. Refuted.	The Negative argument is more strongly supported by the Act itself, and the rejection of the amendments further reinforces this position. Refuted. ✓
Premier Daniel Andrews of Victoria in Australia sold the rights to water to China. **(ID: 282; Not Enough Evidence)**	Evidence shows the premier does not intend to change his approach to China, which implies no current plans to sell water rights. Refuted. ✗	Agents 1 and 2 found no evidence supporting the claim. Agent 3 noted the Premier’s denial to change his approach to China. Refuted. ✗	The Premier fostered an environment conducive to foreign ownership of water rights; China holds significant water entitlements. Supported.	The Premier promoted business cooperation, not water rights deals. China’s ownership does not equate to approved sales. Refuted.	Arguments from both sides rely more on interpretation and inference than on concrete evidence of a transaction. Not Enough Evidence. ✓

Table 2: Case study. Responses are condensed for brevity while preserving original semantics. Erroneous judgments are in red, accurate judgments are in green, and adversarial arguments in DebateCV are in blue.

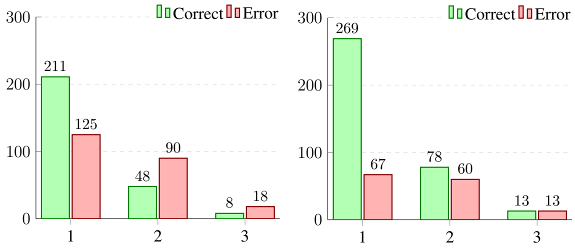
5.3 Further Analysis

We analyze the following key aspects of DebateCV.

What are the benefits of our debate-driven claim verification? To address this, we conduct a case study comparing the claim verification results of the zero-shot DebateCV w/o D-SFT & D-DPO (referred to here simply as DebateCV) with those of the single-agent RAG-COT and multi-agent Majority baselines. We select three claims from the evaluation set, each representing one of the possible verdict categories: Supported (Claim 31), Refuted (Claim 99), and Not Enough Evidence (Claim 282).[6] As shown in Table 2, RAG-COT makes three representative types of errors, while DebateCV’s adversarial debate mechanism exposes these errors and leads to accurate verdicts: (1) **Misinterpretation of claim and/or evidence**: In Claim 31, the RAG-COT misinterprets the date Barrett assumed office (Oct. 27) as contradicting her confirmation date (Oct. 26). In DebateCV, the Debaters propose different interpretations of the term “confirmed”. The Moderator resolves this ambiguity by clarifying that the claim pertains to the confirmation itself, not the assumption of duty, supporting the claim accurately. (2) **Undervaluing direct evidence**: In Claim 99, the baselines support the claim based on the rejection of an amendment but overlook (RAG-COT) or underemphasize (Majority) the explicit language in the Act that mandates medical care for babies born alive. The Nega-

[6] We provide ten additional examples in Appendix E.

tive Debater in DebateCV brings forward this direct evidence, which the baseline and the Affirmative Debater had missed. The Moderator prioritizes the stronger evidence and correctly refutes the claim. (3) **Over-reliance on indirect evidence**: In Claim 282, the RAG-COT refutes the claim based on a subjective inference drawn from an indirect evidence. The Majority finds insufficient evidence to support the claim but fails to recognize that there is also insufficient evidence to refute it. In DebateCV, both Debaters similarly rely on speculative inferences to either support or refute the claim. However, the Moderator accurately identifies the lack of concrete evidence from both sides.



(a) Zero-shot Moderator. (b) Advanced Moderator.

Figure 2: Distribution of debate rounds (x-axis) for correctly and incorrectly predicted claims (y-axis).

What are the benefits of our post-training strategy? To analyze how post-training enhances the Moderator's capabilities, we compare the zero-shot Moderator (DebateCV w/o D-SFT & D-DPO) with the Advanced Moderator (DebateCV) using retrieved evidence, with Llama-3.1 as the base model.

As shown in Fig. 2, the zero-shot Moderator achieves most correct verdicts within the first debate round. However, as debates progress to subsequent rounds, its error rate increases. This may be because claims that requiring extended debate are inherently more complex. Besides, LLM performance is observed to degrade in multi-turn conversations in various tasks (Laban et al., 2025; Li et al., 2023b,c). In contrast, the Advanced Moderator maintains consistently higher accuracy across all rounds, which indicates that our post-training improves the Moderator's ability to make accurate decisions for both straightforward claims resolvable in a single round and complex ones.

Additionally, we observe that the zero-shot Moderator tends to issue neutral verdicts, i.e., Not Enough Evidence (NEE) and Conflicting Evidence/Cherry-Picking (CEC), even when sufficient evidence exists to support or refute a claim. This behavior can arise from conformity bias in LLM agents (Weng et al., 2025; Zhang et al., 2024), where the Moderator leans toward agreeing with both Debaters to avoid definitive judgments that contradict either side. As shown in Table 3, while our debate-driven approach improves zero-shot claim verification compared to the RAG-COT baseline, it also increases conformity, with false positive rates (FPRs) of 9.4% (NEE) and 26.2% (CEC). After applying D-SFT and D-DPO post-training, these rates drop to 2.0% (NEE) and 1.5% (CEC), demonstrating the effectiveness of our training strategy in reducing unwarranted neutrality.

Methods	**Acc.**	**FPR ↓**	
		NEE	CEC
RAG-COT	45.6	4.5%	12.6%
DebateCV	**72.0**	**2.0%**	**1.5%**
DebateCV w/o D-SFT & D-DPO	53.4	9.4%	26.2%

Table 3: False positive rates for neutral verdicts.

How is the computational cost of our framework? Table 4 shows that DebateCV incurs a marginal cost increase ($0.0149 per claim) over RAG-COT, with detailed cost calculations available in Appendix D. Given the importance of accuracy in fact-checking, this extra expense is justified.

Methods	**Input**		**Output**		**Total Cost ($) ↓**
	Tokens	Cost ($)	Tokens	Cost ($)	
RAG-COT	809.19	0.0020	423.95	0.0042	**0.0062**
Majority	4951.14	0.0124	1611.01	0.0161	0.0284
DebateCV	7040.71	0.0089	1816.79	0.0122	0.0211

Table 4: Computational cost per claim verification.

6 Conclusion

In this work, we introduce DebateCV, the first claim verification framework that adopts a debate-driven methodology inspired by human fact-checking practices, using multiple LLM agents. We further propose a novel debate data synthesize method combined with a post-training strategy to improve the framework for better debate-driven claim verification. Our approach enhances evidence analysis during claim verification, resulting in improved accuracy across varying levels of evidence sufficiency. Such advancements improve digital literacy and combat misinformation, which we hope will contribute to a more trustworthy world.

Limitations

The scope of our work is limited to claim verification, with evidence retrieval remaining unaddressed. Nevertheless, we observe a significant influence of the evidence condition, as methods utilizing golden evidence outperform those relying on retrieved evidence from real-world sources. A promising direction for future research can involve extending the DebateCV framework to support iterative evidence retrieval, thereby jointly optimizing both evidence retrieval and claim verification through debate-driven evidence analysis.

Besides, while our proposed methods demonstrate strong performance, they are proposed to consider English-only claims and evidences, constraining the generalization of our results and findings. Misinformation is a global issue that crosses language barriers, highlighting the need for future research to tackle claims in other languages.

In addition, we acknowledge that our proposed debate-driven claim verification may entail higher computational costs than an single agent approach, as the agents must engage in multiple rounds of interaction before reaching a verdict. As detailed in Appendix D, our analysis shows that the DebateCV method costs an additional $0.00149 per claim verification compared to the single-agent RAG approach. However, given the critical importance of accuracy in fact-checking tasks, we believe this additional cost is justified. Future research could explore ways to improve cost efficiency.

References

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proc. of NAACL*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proc. of ICML*.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. 2024. Acc-debate: An actor-critic approach to multi-agent debate. In *Proc. of ICLR*.

Lucas Graves. 2013. *Deciding what's true: Fact-checking journalism and the new ecology of news*. Ph.D. thesis, Columbia University.

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

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. In *Proc. of ICLR*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-rank adaptation of large language models. In *Proc. of ICLR*.

Korir Nancy Jeptoo and Chengjie Sun. 2024. Enhancing fake news detection with large language models through multi-agent debates. In *Proc. of NLPCC*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive LLMs leads to more truthful answers. In *Proc. of ICML*.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can LLMs produce faithful explanations for fact-checking? Towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proc. of ICWSM*.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large language model society. In *Proc. of NeurIPS*.

Ming Li, Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2024a. Can LLMs speak for diverse people? Tuning LLMs via debate to generate controllable controversial statements. In *Proc. of ACL*.

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024b. MCFEND: A multi-source benchmark dataset for chinese fake news detection. In *Proc. of WWW*.

Yupeng Li, Haorui He, Shaonan Wang, Francis CM Lau, and Yunya Song. 2023b. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*, 10(6):3031–3042.

Yupeng Li, Dacheng Wen, Haorui He, Jianxiong Guo, Xuan Ning, and Francis CM Lau. 2023c. Contextual target-specific stance detection on twitter: Dataset and method. In *Proc. of ICDM*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proc. of EMNLP*.

Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See-Kiong Ng, and Tat-Seng Chua. 2025. Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. *arXiv preprint arXiv:2502.17924*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. In *Proc. of NeurIPS*.

Meta. 2023. Llama: Open and efficient foundation language models.

OpenAI. 2024. ChatGPT-4o. https://chat.openai.com/. [Online; accessed 15-October-2024].

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Proc. of NeurIPS*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. In *Proc. of NeurIPS*.

Mark Rothermel, Tobias Braun, Marcus Rohrbach, and Anna Rohrbach. 2024. InFact: A strong baseline for automated fact-checking. In *Proc. of FEVER Workshop*.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024a. The automated verification of textual claims (AVeriTeC) shared task. In *Proc. of FEVER Workshop*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024b. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Proc. of NeurIPS*.

Vighnesh Subramaniam, Antonio Torralba, and Shuang Li. 2024. Debategpt: Fine-tuning large language models with multi-agent debate supervision. https://openreview.net/forum?id=ChNy95ovpF.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, Zhenhan Dai, Yifeng Xie, Yihan Cao, Lichao Sun, Pan Zhou, Lifang He, Hechang Chen, Yu Zhang, Qingsong Wen, Tianming Liu, Neil Zhenqiang Gong, Jiliang Tang, Caiming Xiong, Heng Ji, Philip S. Yu, and Jianfeng Gao. 2025. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*.

Herbert Ullrich, Tomáš Mlynář, and Jan Drchal. 2024. AIC CTU system at AVeriTeC: Re-framing automated fact-checking as a simple RAG task. In *Proc. of FEVER Workshop*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proc. of ACL*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proc. of WWW*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. In *Proc. of ICLR*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversation. In *Proc. of ICLR Workshop on LLM Agents*.

Yuzhou Yang, Yangming Zhou, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2024. Search, examine and early-termination: Fake news detection with annotation-free evidences. In *Proc. of ECAI*.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2024. HerO at AVeriTeC: The herd of open large language models for verifying real-world claims. In *Proc. of FEVER Workshop*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proc. or ACL*.

A Definitions of the Verdict Categories

Similar to (Schlichtkrull et al., 2024b; Rothermel et al., 2024; Yoon et al., 2024; Ullrich et al., 2024), we consider four categories of verdicts for the claims. Below is a summary of the definitions of these verdicts.

1. **Supported**: The claim is supported by the arguments and evidence presented.

2. **Refuted**: The claim is contradicted by the arguments and evidence presented.

3. **Not Enough Evidence**: The presented evidence is not enough to support or refute the

claim. This category applies when the evidence either explicitly indicates that relevant evidence cannot be found or leaves certain aspects of the claim neither supported nor refuted.

4. **Conflicting Evidence/Cherry-Picking**: The claim is misleading due to conflicting evidence or cherry-picking, but is not explicitly refuted. This category includes cases such as cherry-picking (selectively presenting evidence to misrepresent truth), true-but-misleading (e.g., “`Alice has never lost an election`” when Alice has only ever run unopposed), and instances where contradictory evidence is found.

A more detailed description of these categories and additional examples can be found in (Schlichtkrull et al., 2024b).

B Prompts

This section presents the carefully crafted prompts utilized in our work to direct the behavior of the LLM agents, namely the Debaters, Moderator, and Corrector.

B.1 Pre-debate Configuration Prompts

In pre-debate configuration, meta-prompts are provided to both Debaters and Moderator to instruct their actions throughout the multi-round debate.

The meta prompt for *Debaters* is as follows:

> You are a Debater in a fact-checking scenario. Your task is to verify the accuracy of a specific claim by either affirming or refuting it based on the provided credible evidence.
> **Your Role**:
> - Defend your assigned position using credible evidence.
> - Critically analyze and challenge the opposing argument respectfully and factually.
>
> **Objective**:
> - Determine the truthfulness of the claim.
> - Provide compelling evidence to support your stance.
> - Address and critique the opposing evidence to strengthen your position, while being concise.
>
> **Claim**: [CLAIM]
> **Evidence Set**: [EVIDENCE_SET]

The meta prompt for *Moderator* is as follows:

> You are the Moderator in a fact-checking debate where two Debaters examine the truthfulness of a given claim by presenting credible supporting or opposing evidence. Your goal is to facilitate a fact-based evaluation of the claim, ensuring each side maintains a strong, evidence-backed stance.
> **Responsibilities**:
> - Guide each debate round, ensuring arguments remain evidence-based.
> - Assess the relevance and strength of the credible evidence presented by both sides.
> - Determine if further rounds are essential based on the new insights provided. End the debate if both sides are just repeating their previous arguments without bringing new insights.
>
> **Claim**: [CLAIM]
> **Evidence Set**: [EVIDENCE_SET]
> **Verdict Criteria**:
> - Supported: The claim is fully supported by the credible evidence provided.
> - Refuted: The claim is directly contradicted by the credible evidence.
> - Not Enough Evidence: Insufficient credible evidence to confirm or refute the claim.
> - Conflicting Evidence/Cherry-picking: The claim is misleading due to conflicting evidence or selective information.

B.2 In-debate Process Prompts

After the pre-debate configuration, we use the following prompt to instruct the *Affirmative Debater* to initialize the debate.

> **Claim**: [CLAIM]
> You are affirming the veracity of the claim. Use the credible evidence provided to construct a compelling argument supporting the claim.
> **Approach**:
> - Decompose the claim into its core components and clearly articulate its meaning.
> - Present the most relevant evidence to substantiate your position, citing each piece as (Content of evidence, Source URL).
> - Critically analyze and refute the opposition’s points using credible evidence (this can be omitted in the first round).

In rebuttal, the *Negative Debater* is guided to present counterargument using the following prompt.

> **Previous Argument (Affirmative)**: [AFFIRMATIVE_ARGUMENT]
> You are refuting the claim, arguing that it is false. Use the credible evidence provided to construct a compelling argument against the claim.
> **Approach**:
> - Present the most relevant evidence to support your position, citing each piece as (Content of evidence, Source URL).
> - Critically analyze and counter the opposition's points using credible evidence.

At the end of each round, the *Moderator* oversees the debate. Its prompt is as follows:

> Round [ROUND_NUMBER] of the fact-checking debate has concluded.
> **Affirmative**: [AFFIRMATIVE ARGUMENT]
> **Negative**: [NEGATIVE ARGUMENT]
> As the Moderator, evaluate each side's arguments by examining the relevance and sufficiency of the credible evidence presented, while being concise.
> **Steps to Follow**:
> 1. Summarize the main new insights obtained from this round compared to previous rounds.
> 2. Note any missing evidence or arguments in either side's case.
> 3. Assess if further debate is necessary or if the arguments are repeating previous points without adding substantial new information.
> 4. Conclusion:
> -If a clear verdict is supported or no need for further debate: Provide justification for this outcome; Select one of the following Verdict labels: "Supported", "Refuted", "Not Enough Evidence", or "Conflicting Evidence/Cherry-picking"; Set "Proceeding Necessity" to "No".
> -If further debate is essential: Indicate why additional rounds are necessary; Set "Proceeding Necessity" to "Yes"; Provide a "Justification for Proceeding" outlining the rationale for needing further evidence; Leave "Justification for Verdict" and "Verdict" blank.
> Output your findings in JSON format: {"Primary Insight": "...", "Evidence Gaps": "...", "Justification for Proceeding": "...", "Proceeding Necessity": "Yes or No", "Justification for Verdict": "...", "Verdict": "Supported", "Refuted", "Not Enough Evidence", or "Conflicting Evidence/Cherry-picking" }

If the Moderator opts to continue the debate, the following interaction prompt will be provided to the *Affirmative Debater*.

> **Opposition Argument**: [OPPOSITION ARGUMENT]
> Do you agree with this perspective? Provide your response explaining your reasoning, supporting evidence, and highlighting any weaknesses in the opposition's points.

If the Moderator determines that the debate between the two Debaters has converged, the debate will be terminated. In cases where the debate lasts for more than three rounds, the *Moderator* will receive the final prompt as follows to output a verdict and generate a corresponding justification.

> Affirmative: [AFFIRMATIVE_ARGUMENT]
> Negative: [NEGATIVE_ARGUMENT]
> Summarize the primary insights gathered throughout the entire debate concisely.
> After reviewing both sides' arguments on the claim: [CLAIM]
> Select a Verdict from the following labels based on the credible evidence: "Supported", "Refuted", "Not Enough Evidence", or "Conflicting Evidence/Cherry-picking".
> Provide a step-by-step justification for your choice, then present your conclusion in JSON format: { "Justification for Verdict": "...", "Verdict": "Supported", "Refuted", "Not Enough Evidence", or "Conflicting Evidence/Cherry-picking" }

B.3 Corrector's Prompt

This prompt is used to instruct the *Corrector* to generate a justification explaining how the ground-truth verdict can be derived from the debate.

> Debate Recoding: [DEBATE_RECORDING]
> Primary Insights: [Primary_Insights]
> Task: Please provide the justification for your verdict that this claim is [GT_VERDICT] based on the debate context.
> The output should be in JSON format. { "Justification for Verdict": "..." }

C Implementation Details of the Baselines

This appendix provides implementation details of the baselines used in our experiments.

In our evaluation, RAG-COT, Majority, Cooperation, and DebateCV w/o D-SFT & D-DPO utilize both the proprietary model GPT-4o (OpenAI, 2024) and the open-source Llama-3.1-8B-Instruct (Meta, 2023) (referred to as Llama-3.1) to ensure comprehensive evaluation across different LLMs, while post-trained baselines, RAG-SFT, DebateCV, and its w/o D-DPO variant, exclusively employ Llama-3.1 as the backbone due to GPT-4o's inaccessibility for fine-tuning. All post-trained methods leverage LoRA (Hu et al., 2021), a parameter-efficient technique that minimizes computational overhead.

For RAG-SFT, we adopt the open-source Llama-3.1 checkpoint provided by Yoon et al. (2024).[7] The Majority baseline employs the voting prompt from Lin et al. (2025)'s official implementation,[8] while the Cooperation baseline adapts the CAMEL framework for claim verification,[9] ensuring alignment with its official repository.

For all the LLMs, the maximum number of generated tokens, temperature, and top_p are set to 512, 0.7, and 1.0, respectively. In the post-trained methods involving training, we use AdamW optimizer with a learning rate of 2×10^{-5} for 2 epochs. For LoRA, we set the rank to 128 and alpha to 256. These hyper-parameters follow the settings used in Yoon et al. (2024) to ensure a fair and direct comparison with their results. All experiments involving Llama-3.1 were conducted on a single 40GB NVIDIA A100 GPU. Proprietary models such as GPT-4o and GPT-4o-mini were accessed via OpenAI's API.

D Detailed Computational Cost Analysis

Methods	Input		Output		Total Cost ($)
	Tokens	Cost ($)	Tokens	Cost ($)	
RAG-COT	809.19	0.0020	423.95	0.0042	0.0062
Majority	4951.14	0.0124	1611.01	0.0161	0.0284
DebateCV					
Debaters	3685.13	0.0005	1526.01	0.0009	0.0014
Moderator	3355.58	0.0084	290.78	0.0113	0.0197
Total	7040.71	0.0089	1816.79	0.0122	0.0211

Table 5: Computational cost per claim verification.

In this section, we compare the inference costs of our zero-shot DebateCV w/o D-SFT & D-DPO (referred to here simply as DebateCV) against the single-agent RAG-COT and multi-agent Majority baselines under the realistic retrieved evidence setup based on the latest pricing from the OpenAI API. In DebateCV, the Debaters are powered by GPT-4o-mini (Input: \$0.15 per 1M tokens, Output: \$0.60 per 1M tokens), while the Moderator utilizes GPT-4o (Input: \$2.5 per 1M tokens, Output: \$10.00 per 1M tokens). Table 5 summarizes the average token usage and associated costs for verifying a single claim. The results show that DebateCV incurs a slightly higher cost compared to the single-agent RAG approach, with an average difference of \$0.0149 per claim verification (\$0.0211 vs. \$0.0062). As seen in other downstream applications based on multi-agent debate frameworks (Du et al., 2024; Liang et al., 2024; Lan et al., 2024), this additional cost is acceptable given the significance of accuracy in the fact-checking task. For high-stakes applications such as misinformation detection or decision-support systems, even small improvements in accuracy can have a meaningful impact in the real world.

E Examples of DebateCV Outputs

To demonstrate how DebateCV performs claim verification, we provide ten examples in Table 6, in addition to the case study detailed in Section 5.3. Each example consists of an input claim, followed by arguments presented by the Affirmative and Negative Debaters, and concludes with the Moderator's final decision. We use the following prompt to condense agent responses across all debate rounds for brevity, using GPT-4o-mini:

> This a claim verification debate that involves two Debaters (Affirmative and Negative) and a Moderator. The claim under consideration is: [CLAIM]
> The full transcript of the debate is as follows: [DEBATE_RECORD]
> Summarize the arguments presented by both sides and the Moderator's final decision. Write each summary in the first person from the perspective of each Debater and the Moderator.

[7] https://huggingface.co/humane-lab/Meta-Llama-3.1-8B-HerO. Licensed under CC BY-NC 4.0.

[8] https://github.com/DanielLin97/FACT-AUDIT

[9] https://github.com/camel-ai/camel

Claim (ID; Label)	DebateCV		
	Affirmative Debater	Negative Debater	Moderator
UNESCO declared Nadar community as the most ancient race in the world. **(ID: 3; Refuted)**	The Nadar community has a rich historical presence and cultural significance. Some scholars recognize their deep roots. Supported.	No direct recognition from UNESCO supporting this claim. Many communities have ancient lineages, and cultural prominence does not equate to being the "most ancient." Refuted.	While the Affirmative highlights historical importance, it lacks official UNESCO endorsement. The Negative argument effectively refutes the lack of formal recognition. Refuted.
The United States and its Western allies have been using their media outlets to publish articles based on fabricated information under allegations of non-compliance with the Chemical Weapons Convention. **(ID: 5; Refuted)**	There are accusations from multiple sources—including pro-Kremlin and Western outlets—that narratives around Syria's compliance were manipulated. OPCW reports have also been contested. Supported.	OPCW investigations consistently raised concerns about Syria's compliance, supported by verified findings. Reports of chemical use are corroborated by independent missions. Allegations are based on real evidence. Refuted.	The Negative side provided stronger and more concrete evidence, including verified OPCW findings. Affirmative claims lacked proof of systemic fabrication. Refuted.
"Why should you pay more taxes than Donald Trump pays?" Biden cited $750 paid by Trump in 2016 and 2017, and quoted Trump saying, "Because I'm smart. I know how to game the system." **(ID: 7; Supported)**	Multiple reputable sources (NYT, NPR, Forbes) confirm Trump paid only $750 in federal income tax during those years. His own quote suggests he exploited legal loopholes. Supported.	Breitbart disputes the NYT report, suggesting it misrepresents Trump's filings. Not all wealthy individuals avoid taxes, but many legally reduce liabilities through established strategies. The claim oversimplifies tax complexities. Refuted.	Despite counterarguments, the Affirmative side presented credible and consistent evidence from multiple reputable sources. The Negative side failed to refute the core facts. Supported.
Joe Biden wants to ban fracking. **(ID: 23; Refuted)**	Biden supports banning new oil and gas permits on public lands, which limits fracking. While he hasn't explicitly called, his campaign indicated restrictions that align with a de facto ban. Supported.	Biden has explicitly stated he does not support a complete ban on fracking. He aims to restrict new permits on public lands, not eliminate existing operations. Campaign statements clarify this nuance. Refuted.	The Negative correctly identifies Biden's nuanced stance. His policies target new permits, not an outright ban. Refuted.
The New York Post endorsed Donald Trump's bid for re-election. **(ID: 32; Supported)**	The New York Post editorial board explicitly endorsed Trump in the 2020 election, despite some policy criticisms. Endorsements can reflect strategic alignment rather than full agreement. Supported.	While the Post has conservative leanings, its endorsement included reservations and critiques, suggesting a complex rather than clear-cut stance. Refuted.	The Affirmative presents direct evidence of the endorsement, while the Negative raises contextual concerns without disproving it. Supported.
COVID-19 pandemic pre-planned with the help from Bill Gates. **(ID: 35; Refuted)**	Gates long warned of pandemics and participated in Event 201. These suggest foreknowledge and planning, fueling public perception of involvement. Supported.	Gates' advocacy was for preparedness, not prediction. Event 201 was a simulation, not a blueprint. Conspiracy theories lack factual basis. Refuted.	Affirmative relies on correlation and speculation, while the Negative clearly distinguishes between warning and orchestration. Refuted.
Breitbart News reports that the daughter of Delaware Democratic Senator Chris Coons and seven other underage girls were featured on Hunter Biden's laptop. **(ID: 37; Refuted)**	Allegations from political figures like Lauren Witzke suggest there may be merit to the claim. Lack of denial implies some level of credibility. Supported.	No credible source confirms Breitbart reported this. Fact-checks disprove child pornography claims. Unsupported allegations should not be treated as fact. Refuted.	Affirmative relies on unverified allegations, while the Negative shows absence of actual reporting or evidence. Refuted.
The umbilical cord should not be cut until 1 hour after birth or the baby will not have enough blood right after birth. **(ID: 46; Refuted)**	ACOG recommends delayed clamping (30–60 sec) to improve hemoglobin and iron levels. A 1-hour delay may offer additional benefits. Supported.	ACOG supports only 30–60 seconds of delay. Longer delay lacks evidence and may not be practical. Refuted.	Delayed clamping is beneficial, but notes no support for 1-hour delay. Medical guidelines back shorter window. Refuted.
People who do not vote for the Bharatiya Janata Party (BJP) in the 2020 elections will not get the COVID vaccine free of cost. **(ID: 51; Refuted)**	BJP manifesto promised free vaccines, but critics feared this might be tied to voter loyalty. Political context implied exclusion. Supported.	Official statements clarified vaccines would be available to all priority groups regardless of political affiliation. Refuted.	No evidence linking eligibility to voting behavior. Government immunization programs apply universally. Refuted.
Officer who wore Trump 2020 mask to polls to face disciplinary action. **(ID: 63; Supported)**	A Miami police officer was photographed wearing a 'Trump 2020' mask at a polling location. NBC News and Miami PD confirm disciplinary action is coming. Mayor Suarez affirmed consequences. Supported.	While the officer likely wore the mask and may face discipline, the term 'disciplinary action' is vague. Free speech rights and political expression complicate the issue. Refuted.	Multiple evidence confirms both the mask-wearing and disciplinary action. The Negative raises valid questions but doesn't contradict the central claim. Supported.

Table 6: Examples of DebateCV outputs for claim verification. The accurate judgments are in green.