Debating the Unspoken: Role-Anchored Multi-Agent Reasoning for Half-Truth Detection

Anonymous ACL submission

Abstract

Half-truths, claims that are factually correct yet misleading due to omitted context, remain a blind spot for fact verification systems focused on explicit falsehoods. Addressing such omission-based manipulation requires reasoning not only about what is said, but also about what is left unsaid. We propose RADAR, a role-anchored multi-agent debate framework for omission-aware fact verification under realistic, noisy retrieval. RADAR assigns complementary roles to a *Politician* and a *Scientist*, who reason adversarially over shared retrieved evidence, moderated by a neutral *Judge*. A dual-threshold early termination controller adaptively decides when sufficient reasoning has been reached to issue a verdict. Experiments show that RADAR consistently outperforms strong single- and multi-agent baselines across datasets and backbones, improving omission detection accuracy while reducing reasoning cost. These results demonstrate that role-anchored, retrieval-grounded debate with adaptive control is an effective and scalable framework for uncovering missing context in fact verification.

1 Introduction

The rapid spread of online misinformation has intensified the demand for scalable and reliable fact verification systems (OpenAI et al., 2024; Hazra and Majumder, 2024; Estornell et al., 2020). While large language models (LLMs) have made notable progress in detecting explicit falsehoods, a subtler and more pervasive form of manipulation remains difficult to identify: the *half-truth*. Half-truths are claims that are factually correct in isolation but misleading due to omitted context. For instance, stating that a politician reduced national debt by 15% may be accurate, yet misleading if it conceals that the reduction followed a prior 20% increase during the same term. Such selective framing distorts interpretation without fabricating facts, posing unique challenges for both human and automated verifiers (Singamsetty et al., 2023; Shukla and Bhattacharyya, 2025).

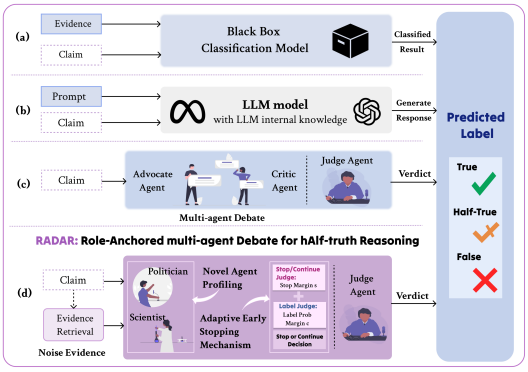


Figure 1: Overview of fact verification paradigms: (a) encoder-based classifiers over claim–evidence pairs, (b) prompt-guided LLMs, (c) multi-agent debate with fixed *pro/con* roles and internal knowledge, and (d) our proposed RADAR, which performs omission-aware reasoning through evidence-grounded role debate.

Detecting half-truths differs fundamentally from standard fact verification. Rather than identifying explicit contradictions, omission-aware verification must determine whether the available evidence is insufficient to justify a claim's implied conclusion, requiring models to reason over latent intermediate assumptions rather than surface support signals. This challenge is exacerbated under realistic retrieval settings, where evidence is noisy and incomplete. Even when retrieved passages appear supportive, they may omit critical background information, causing systems to misclassify omission-based claims as true.

Most existing fact verification pipelines are not designed for this setting. As shown in Figure 1, single-agent approaches perform one-pass reasoning over retrieved evidence and are prone to error propagation when key context is missing. Multi-agent debate (MAD) frameworks (Liu et al., 2025; Han et al., 2025) introduce reasoning diversity

through argument exchange, but their fixed *pro/con* roles and focus on explicit contradiction make them ill-suited for omission reasoning, where the core issue is missing context rather than opposing claims. Recent work TRACER (Tang et al., 2025) has begun to explicitly study omission-aware verification, but assumes access to gold evidence and operates as a single-agent pipeline without adaptive reasoning, limiting its applicability under realistic retrieval conditions.

In this paper, we propose RADAR (**R**ole-**A**nchored multi-agent **D**ebate for h**A**lf-t**R**uth reasoning), a framework for omission-aware fact verification under incomplete retrieval. Unlike conventional debate systems with fixed *pro/con* stances or internal knowledge reliance, RADAR formulates verification as a structured debate among agents with realistic expertise roles, all constrained to reason over retrieved evidence. A *Politician* agent constructs a coherent narrative supporting the claim based on available evidence, while a *Scientist* agent probes for missing, weak, or selectively presented context. Their interaction is moderated by a neutral *Judge*, which synthesizes the debate and issues the final verdict.

To enhance efficiency, RADAR incorporates a dual-threshold early termination controller that monitors both the judge's stop margin and verdict confidence. The debate concludes automatically once sufficient information has been exchanged, preventing redundant interactions while preserving reasoning depth. Through this adaptive mechanism, RADAR balances interpretability, accuracy, and computational cost, making omission reasoning more scalable for real-world verification.

Our contributions are as follows:

- We address the underexplored challenge of half-truth detection under realistic noisy retrieval, where gold evidence is unavailable.
- We propose RADAR, a role-anchored, retrieval-grounded multi-agent framework that surfaces omission-based half-truths through complementary expertise debate, with adaptive early stopping for efficiency.
- We demonstrate across datasets and backbone LLMs that RADAR consistently improves both accuracy and efficiency over strong single-agent and multi-agent baselines.

2 Related Work

2.1 Fact Verification and Half-Truths

Fact verification is commonly formulated as determining whether a claim is supported or refuted given retrieved evidence (Thorne et al., 2018; Alam et al., 2021; Zhang and Gao, 2023; Wu et al., 2025), with numerous systems exploring retrieval, reasoning, and explanation (Chen et al., 2022a; Sriram et al., 2024; Yue et al., 2024; Wang et al., 2025). CLIMINATOR (Leippold et al., 2024) studies LLM-based fact verification in the climate domain. CLAIMDECOMP (Chen et al., 2022b) decompose claims into sub-questions, while FIRE (Xie et al., 2025) uses iterative verify-or-search loops to reformulate queries upon failure. While effective for explicit misinformation, these approaches largely assume that misleading claims manifest as direct contradictions with evidence.

Half-truths violate this assumption. Claims may be factually correct with respect to retrieved evidence yet still be misleading due to omitted context. TRACER (Tang et al., 2025) is the first framework to explicitly model this phenomenon through the concept of *critical hidden evidence*, but assumes access to gold evidence annotations and follows a single-agent, static reasoning pipeline. Our work addresses omission detection under noisy retrieval through structured multi-agent interaction.

2.2 Multi-Agent Debate for Reasoning

Multi-agent debate (MAD) has recently emerged as a paradigm for enhancing reasoning and factuality in LLMs (Du et al., 2024; Wei et al., 2022; Liang et al., 2024; Guo et al., 2024; Tillmann, 2025; Liu et al., 2025). Debate-based frameworks such as D2D (Han et al., 2025) and TruEDebate (TED) (Liu et al., 2025) employ multi-stage dialogue between a proponent, skeptic, and judge to reach veracity judgments for fake news. However, most MAD systems are designed for contradiction-oriented verification, where agents argue for or against a claim. Omission-based half-truths differ fundamentally: the key issue may be missing context rather than explicit conflict. RADAR therefore replaces polarity-based debate with complementary reasoning roles and grounds all interaction in retrieved evidence, enabling agents to surface what is absent rather than only dispute what is stated.

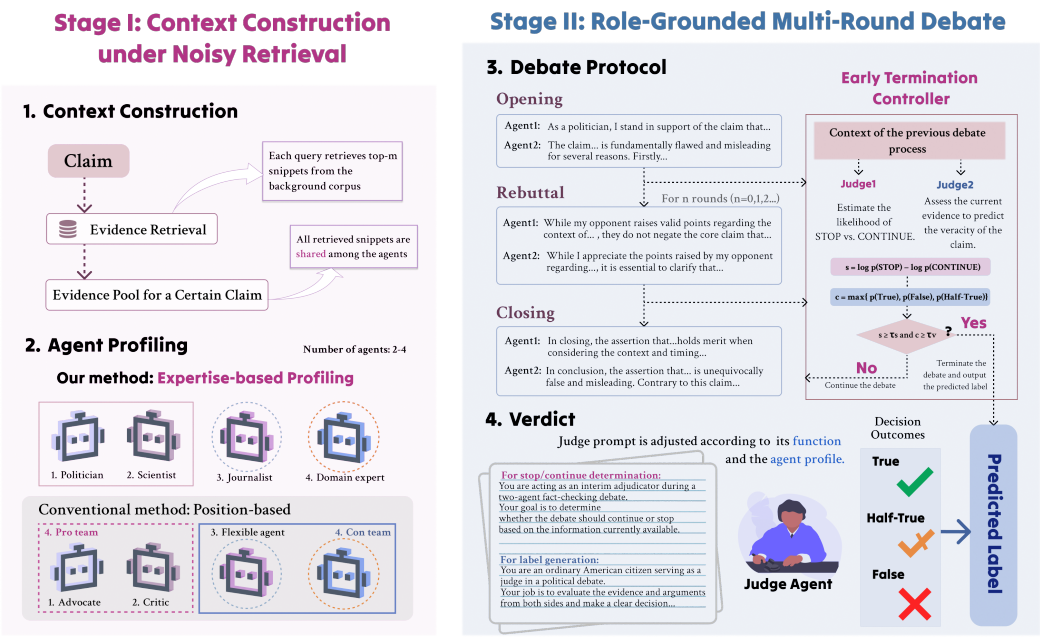


Figure 2: Overview of the RADAR framework for omission-based half-truth detection. The system conducts structured multi-agent debate between expertise-grounded roles over retrieved evidence, equipped with an adaptive early termination controller to uncover missing yet critical context efficiently.

2.3 Bias by Omission in News

Ehrhardt et al. (2021) detect political slant by analyzing entity co-occurrence across outlets. Zhukova et al. (2023) model bias by commission, omission, and source selection through the COSS framework. Other works explore rhetorical manipulation and propaganda framing (Martino et al., 2019; Akhtar et al., 2024; Piskorski et al., 2023) or examine linguistic cues distinguishing factual from deceptive reporting (Rashkin et al., 2017). These studies highlight omission as a key mechanism of bias. Our work formulates omission as a reasoning and verification task, aiming to detect when incomplete context leads to misleading claims.

3 RADAR Framework

3.1 Framework Overview

We present **RADAR** (*Role-Anchored multi-agent Debate for hAlf-truth Reasoning*), a multi-agent framework for detecting omission-based half-truths under realistic retrieval conditions. Unlike prior systems that rely on single-pass classification or assume gold evidence, RADAR performs deliberative reasoning among role-specialized agents who exchange arguments grounded in retrieved evidence. As illustrated in Figure 2, the framework operates in two stages: (1) *Context Construction under Noisy Retrieval*, which builds a shared evidence pool without assuming completeness, and (2) *Role-Grounded Multi-Round Debate*, where agents reason adversarially to identify missing context and assess whether the claim is misleading.

Role	Objective and Reasoning Bias
Politician	Constructs a coherent, evidence-based narrative; tends to emphasize confirmatory context.
Scientist	Examines retrieved evidence for missing or contradicting information; tends to challenge selective framing.
Judge	Evaluates debate sufficiency, coherence, and evidential coverage; decides when to stop and issues the final verdict.

Table 1: Roles and reasoning types in RADAR.

Given an input claim c, RADAR constructs an evidence pool E by retrieving top-ranked text segments from a background corpus. Three agents then participate in the subsequent debate: a *Politician*, a *Scientist*, and a neutral *Judge*, whose roles and reasoning styles are summarized in Table 1. This design transforms verification from stance competition into omission-focused reasoning, with one agent constructing the strongest evidence-backed narrative while the other probes for missing or underrepresented context.

3.2 Stage I: Context Construction under Noisy Retrieval

Omission-based misinformation often remains convincing because relevant facts are selectively pre-

sented while critical context is absent. To evaluate claims under realistic conditions, RADAR first constructs a retrieval-grounded evidence pool that is informative yet imperfect. Rather than assuming complete supporting evidence, the system requires downstream reasoning to operate under uncertainty, where retrieved information may be partial, noisy, or misleading.

Evidence Retrieval. Given a claim c, RADAR retrieves the top-m relevant snippets from a background corpus to form an evidence pool $E = \text{RETRIEVE}(c)$. Because retrieval is inherently imperfect, E may contain irrelevant passages, fragmented facts, or only partially supportive evidence. This setting better reflects real-world fact verification, where systems must reason from incomplete retrieved context. It differs from prior omission-aware work such as TRACER (Tang et al., 2025), which assumes access to gold evidence, and from debate frameworks such as TED (Liu et al., 2025) or D2D (Han et al., 2025), which rely primarily on internal model knowledge.

Shared Grounding. The same evidence pool E is provided to all agents as a common factual basis. During debate, agents are required to justify their arguments using retrieved evidence, which improves transparency and reduces unsupported claims. Since all agents observe identical but incomplete evidence, differences in their conclusions arise from reasoning rather than information asymmetry.

3.3 Stage II: Role-Grounded Multi-Round Debate

Once the shared context is established, RADAR conducts a structured debate in which the Politician and Scientist iteratively exchange arguments while the Judge moderates and evaluates progress.

Debate Protocol. Each debate instance begins with an *opening round*, where the Politician presents an evidence-backed narrative supporting the claim and the Scientist highlights potential omissions or contradictory details. In the *rebuttal round*, both agents are tasked with challenging the opponent's argument by identifying potential flaws or weaknesses. Finally, in the *closing round*, both summarize their stances, after which the Judge synthesizes the debate transcript and evidence to assign one of three labels: *true*, *half-true*, or *false*.

Adaptive Early Stopping. To balance reasoning thoroughness and efficiency, RADAR introduces a dual-threshold *early termination controller* that determines whether to continue or stop after each round. At the end of every round, the Judge model evaluates the accumulated dialogue and computes the *stop margin* as

$$s = p(\text{STOP}) - p(\text{CONTINUE})$$

where $p(\cdot)$ denotes the softmax-normalized probability of the corresponding decision token. The debate proceeds if $s < \tau_s$, and terminates otherwise. To prevent premature stopping on ambiguous cases (especially half-truths), we further require the model's veracity confidence

$$c = \max_{y \in \{\text{True,False,Half-True}\}} p(y)$$

to exceed a second threshold: early termination occurs only when $s \geq \tau_s$ *and* $c \geq \tau_v$. Both thresholds are calibrated on a development set using cached decision scores from a single inference pass. This adaptive control balances reasoning depth with computational cost, allowing RADAR to maintain accuracy while shortening redundant interactions.

Structured Verdict and Interpretability. At the end of the debate, the Judge produces a structured output:

[REASON]: *Concise explanation citing evidence and debate exchanges.*

[VERDICT]: *TRUE / FALSE / HALF-TRUE*

The inclusion of debate transcripts and evidence citations provides transparent reasoning paths that expose which contextual gaps or omissions influenced the final decision. This interpretability distinguishes RADAR from opaque end-to-end verifiers, enabling human fact-checkers to trace and validate each reasoning step.

By integrating retrieval-grounded context, expertise-based roles, and adaptive termination, RADAR transforms omission detection from static classification into a dynamic reasoning process. All prompt templates and implementation details are presented in Appendix A.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate RADAR on the POLITIFACT-HIDDEN benchmark (Tang et al., 2025), which contains ∼15k political claims with

Model	Accu.	$F1_{mc}$	$F1_T$	$F1_{HT}$	$F1_F$
Full Evidence Setting					
CoT	76.3	64.3	52.9	52.8	87.1
CoT + RA	78.5	68.0	56.7	60.2	87.1
HiSS	74.7	61.9	48.1	51.2	86.4
HiSS + RA	77.8	65.7	49.1	61.5	86.5
$\text{RADAR}_{\text{single}}$	68.7	62.4	57.6	51.1	78.6
$\text{RADAR}_{\text{multi}}$	**83.6**	**70.8**	**59.3**	**61.1**	**92.0**
Retrieved Evidence Setting					
FIRE	60.3	46.9	30.7	34.1	75.9
D2D	63.0	50.9	36.3	39.7	76.5
$\text{RADAR}_{\text{single}}$	58.4	51.0	41.2	41.5	70.2
$\text{RADAR}_{\text{multi}}$	**77.7**	**63.3**	**45.6**	**56.5**	**87.6**

Table 2: Main results under full and retrieved evidence settings using `GPT-4o-mini`. "+RA" denotes TRACER-enhanced baseline methods. $\text{RADAR}_{\text{single}}$ and $\text{RADAR}_{\text{multi}}$ share identical architectures and retrieval setups, differing only in whether multi-agent debate is used.

gold evidence and three labels (*true*, *half-true*, *false*), focusing on omission-based reasoning rather than explicit contradiction. The evidence corpus includes 41,952 documents, and we follow the original train/dev/test split (11,994/1,000/2,000). As RADAR is unsupervised, thresholds are tuned on the dev set and results are reported on the test set.

To assess generalization, we further evaluate RADAR on AVERITEC (Schlichtkrull et al., 2023), a real-world claim verification benchmark that retrieves multi-step question-answer evidence from the open web. It features a "Conflicting Evidence/Cherry-picking" label for claims that are technically accurate but misleading due to omitted context, making it a natural testbed for omission-aware verification.

Settings. We evaluate under two conditions. In the *full evidence* setting, models receive all gold evidence per claim. In the *retrieved evidence* setting, we retrieve top-20 passages per claim using a dense retriever (`bge-base-en-v1.5`), introducing noise and incompleteness to test whether RADAR's debate mechanism can recover missing context. Sensitivity to retrieval depths of $top-k \in \{1, 3, 5, 10, 20\}$ is detailed in Appendix D. The same role prompts and debate protocol are used across all datasets and backbones, with no dataset-specific tuning.

To isolate the contribution of multi-agent debate, we include a single-agent variant ($\text{RADAR}_{\text{single}}$) with identical architecture, prompts, and retrieval but without inter-agent interaction; we denote the full system as $\text{RADAR}_{\text{multi}}$.

Baselines. We compare against representative fact verification and debate-based methods:

CoT (Wei et al., 2022): Chain-of-thought prompting over provided evidence.

HiSS (Zhang and Gao, 2023): Hierarchical structured prompting with intermediate reasoning steps.

TRACER (Tang et al., 2025): An omission-aware verification framework that assumes access to complete gold evidence, including hidden context required for intent and assumption inference.

FIRE (Xie et al., 2025): An iterative retrieve-then-verify framework alternating between retrieval and validation.

D2D (Han et al., 2025): A multi-agent debate model for news misinformation detection.

For all baselines, we follow the experimental settings recommended in their original papers to reflect their intended usage. For prompting-based methods such as CoT, we use a standard task instruction with step-by-step reasoning. For framework-specific baselines such as FIRE and D2D, we preserve their original prompting protocols. All shared components (backbone models, retrieval pipeline, evidence pool) are kept identical across methods to isolate the effect of reasoning frameworks.

Evaluation Metrics. We report accuracy and macro-F1 as overall metrics, along with per-class F1 scores (*true*, *half-true*, *false*) to analyze class-level behavior. Performance changes for $\text{RADAR}_{\text{multi}}$ are measured relative to its single-agent variant. Each debate proceeds for up to three rounds, with an adaptive early stopping mechanism.

4.2 Main Results

Table 2 presents the main comparison results.

Performance with Gold Evidence. When complete gold evidence is provided, most baselines achieve competitive performance, reflecting that omission-aware verification becomes easier when all relevant context is available. Nevertheless, $\text{RADAR}_{\text{multi}}$ attains the best overall performance, achieving 83.6 accuracy and 70.8 macro-F1, outperforming strong baselines such as CoT+RA and HiSS+RA by clear margins. Compared to $\text{RADAR}_{\text{single}}$, multi-agent debate yields substantial gains (+14.9 accuracy and +8.4 macro-F1), indi-

Model	Backbone	Accuracy	F1$_{macro}$	F1$_{true}$	F1$_{half\text{-}true}$	F1$_{false}$
FIRE	GPT-4o mini	60.3	46.9	30.7	34.1	75.9
D2D		63.0	50.9	36.3	39.7	76.5
RADAR$_{\text{single}}$		58.4	51.0	41.2	41.5	70.2
RADAR$_{\text{multi}}$		**77.7** ↑ 19.3	**63.3** ↑ 12.3	**45.6** ↑ 4.4	**56.5** ↑ 15.0	**87.6** ↑ 17.4
FIRE	LLaMA3-8B-Instruct	50.4	43.4	30.9	39.7	59.6
D2D		64.1	45.1	21.7	34.1	79.5
RADAR$_{\text{single}}$		57.3	52.0	44.4	44.8	66.6
RADAR$_{\text{multi}}$		**77.3** ↑ 20.0	**63.2** ↑ 11.2	**46.3** ↑ 1.9	**56.7** ↑ 11.9	**86.5** ↑ 19.9
FIRE	Qwen2.5-7B-Instruct	63.9	46.9	29.2	32.1	79.5
D2D		69.0	48.5	22.2	41.2	82.1
RADAR$_{\text{single}}$		71.2	54.9	39.4	41.1	84.3
RADAR$_{\text{multi}}$		**76.7** ↑ 5.5	**63.1** ↑ 8.2	**48.2** ↑ 8.8	**54.9** ↑ 13.8	**86.2** ↑ 1.8

Table 3: Generalization results of strong baselines and RADAR variants across different backbone LLMs under the retrieved evidence setting.

Model	Accu.	F1$_{mc}$	F1$_{T}$	F1$_{HT}$	F1$_{F}$
FIRE	26.59	19.38	7.74	7.31	43.84
D2D	60.17	46.38	45.22	19.36	49.6
RADAR$_{\text{multi}}$	**74.09**	**57.87**	**71.68**	**23.82**	**85.05**

Table 4: Performance comparison of baseline methods and RADAR$_{\text{multi}}$ on AVERITEC.

cating that structured interaction between expertise-anchored agents improves reasoning consistency even when evidence is complete.

Performance with Retrieved Evidence. The retrieved-evidence setting is more realistic and challenging due to noisy and incomplete evidence. RADAR$_{\text{multi}}$ remains robust, achieving 77.7 accuracy and 63.3 macro-F1, substantially outperforming all strong baselines. Relative to its single-agent counterpart, multi-agent debate brings large improvements (+19.3 accuracy and +12.3 macro-F1), demonstrating that RADAR's structured interaction is particularly effective when critical evidence is missing or fragmented.

Per-Class Performance Analysis. Per-class results show that performance gains are concentrated in the *half-true* and *false* categories, where omission reasoning is most critical. Under retrieved evidence, RADAR$_{\text{multi}}$ improves *half-true* F1 by 15 and *false* F1 by 17.4, while also maintaining stronger performance on the *true* class. The pronounced gap between RADAR$_{\text{single}}$ and RADAR$_{\text{multi}}$ on these classes suggests that debate helps scrutinize selectively framed narratives and surface omitted context beyond what single-pass reasoning can achieve. To further examine model behavior near class boundaries, especially the subtle distinction between TRUE and HALF-TRUE, we complement the per-class F1 scores in Table 3 with confusion matrices and class-wise precision/recall analyses in Appendix B.

Generalization Across Backbone LLMs. Table 3 reports results across three backbone LLMs. RADAR$_{\text{multi}}$ consistently achieves strong performance across `GPT-4o-mini`, `LLaMA3-8B-Instruct`, and `Qwen2.5-7B-Instruct`. While absolute performance varies across backbones, the relative improvements over strong baselines and the single-agent variant remain stable, indicating that RADAR's debate mechanism generalizes across architectures.

Generalization Across Fact Verification Datasets. Table 4 reports results on the AVERITEC benchmark under the retrieved-evidence setting. According to the table, RADAR$_{\text{multi}}$ substantially outperforms both FIRE and D2D, with consistent gains across all classes. Overall, these results show that RADAR is not limited to political claims. The same role prompts transfer effectively to AVERITEC without task-specific redesign, suggesting that the gains arise from complementary reasoning dynamics rather than domain-specific priors.

5 Discussions

5.1 Impact of Agent Role Configurations

To assess how role design affects reasoning quality, we compare two configurations: the proposed *expertise-based* pairing (Politician vs. Scientist) and the conventional *position-based* pairing (Advocate vs. Critic). As shown in Figure 3, expertise-based agents consistently outperform position-

Model	Acc.	F1	# Avg. tokens
LLaMA3-8B-Instruct			
D2D	64.1	45.1	2483.3
FIRE	50.4	43.4	**754.5** (30.4%)
RADAR (3 rounds)	77.2	62.3	1834.7 (73.9%)
RADAR (early stop)	**77.3**	**63.2**	1551.4 (62.5%)
Qwen2.5-7B-Instruct			
D2D	69.0	48.5	3686.9
FIRE	63.9	46.9	**1026.2** (27.8%)
RADAR (3 rounds)	76.5	62.6	1751.5 (47.5%)
RADAR (early stop)	**76.7**	**63.1**	1723.7 (46.8%)

Table 5: Effect of adaptive early stopping on performance and efficiency.

Backbone	τ_s	τ_v	Stop@1	@2	@3
LLaMA3-8B	−0.15	0.7	846	265	889
Qwen2.5-7B	0.2	0.9	104	3	1893

Table 6: Thresholds and number of instances stopped at each round for RADAR's early-stopping controller.

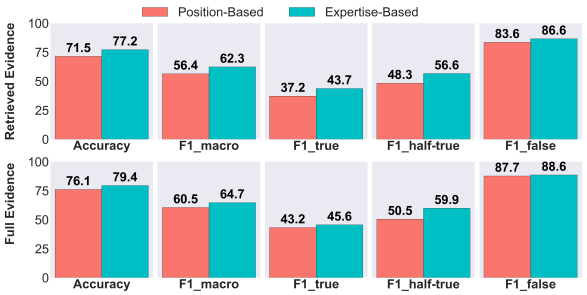
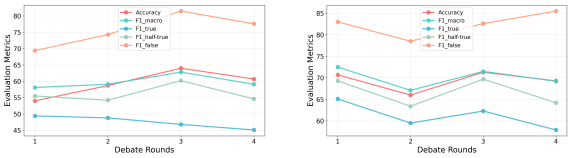


Figure 3: Performance comparison of different agent role configurations under full and retrieved evidence settings.



(a) Retrieved evidence (b) Full evidence

Figure 4: Effect of varying the maximum number of debate rounds using `LLaMA3-8B-Instruct`.

based ones across all settings, with the largest gains observed under noisy retrieval (62.3 vs. 56.4 macro-F1). These results suggest that domain-aligned personas encourage contextually grounded and rhetorically realistic reasoning, exposing omissions that generic pro/con roles often overlook.

5.2 Effect of Early Stopping on Efficiency

Table 5 compares RADAR with adaptive early stopping against D2D, FIRE, and the fixed three-round variant. Across backbones, early stopping preserves or slightly improves performance while reducing generation cost. On `LLaMA3-8B-Instruct`, early stopping achieves comparable accuracy (77.3 vs. 77.2) and higher macro-F1 (63.2 vs. 62.3) while reducing average token usage by ∼15%. On `Qwen2.5-7B-Instruct`, it similarly yields small gains with nearly identical token usage. In both cases, RADAR substantially outperforms D2D in accuracy and macro-F1.

As shown in Table 6, many instances terminate after one or two rounds, indicating that the Judge stops debate once sufficient reasoning confidence is reached while harder cases proceed further. Although optimal thresholds vary across backbones, calibration requires only a single inference pass on the development set with offline grid search over cached scores. Sensitivity analysis in Appendix C confirms stability across broad threshold ranges.

5.3 Effect of Maximum Debate Rounds

We investigate how the maximum number of debate rounds affects performance on a sampled test set of 150 examples (50 per class) using `LLaMA3-8B-Instruct`. As shown in Figure 4, under retrieved evidence, performance improves from one to three rounds as extra interaction helps agents challenge partial evidence and uncover omitted context, with marginal gains beyond that. Under full evidence, the effect is less consistent since most relevant context is already available. These results further motivate RADAR's adaptive early-stopping mechanism, which allocates debate rounds dynamically based on claim difficulty.

5.4 Effect of Number of Agents

To assess the impact of scaling up the number of agents, we vary the number of debating agents using the same 150 set with fixed backbone `LLaMA3-8B-Instruct` and expertise-based roles. As shown in Table 8, performance peaks with two agents (64.0 accuracy, 62.8 macro-F1), confirming that the Politician–Scientist pairing provides complementary yet focused reasoning. Adding more agents yields diminishing returns and lower $F1_{true}$ (dropping to 30.0 with three agents), as additional perspectives often introduce redundant or conflicting arguments. A minimal yet diverse pair thus offers efficient and stable omission reasoning.

Single Agent	Position-Based (Adv. vs Crit.)	Expertise-Based (Pol. vs Sci.)
Verdict: True Klitschko sold his medal in 2012. Funds supported his foundation. The buyer returned it. No evidence disputes the claim.	**Verdict: True** *Pro Opening:* Klitschko sold his medal in 2012. Proceeds went to help Ukrainian children. *Con Opening:* The sale happened in 2012. It's not tied to the 2022 war. *Pro Rebuttal:* The act still helped children and shows lasting support. *Con Rebuttal:* The timing and intent are unrelated to the current crisis. *Pro Closing:* The sale reflects long-term commitment. *Con Closing:* The event came before the war. Linking them misleads. *Judge:* TRUE. The claim is factually accurate.	**Verdict: Half-True** *Politician Opening:* Klitschko auctioned his medal in 2012 to help children. *Scientist Opening:* The sale was real but not linked to the 2022 war. *Politician Rebuttal:* The act still shows support for Ukrainian children. *Scientist Rebuttal:* The claim misrepresents time and cause. *Politician Closing:* The act is real, but how it's framed changes how people remember it. *Scientist Closing:* It was not a wartime gesture. Context is key. *Judge:* HALF-TRUE. The sale was charitable, but the implied timing misleads.

Table 7: Reasoning process for a temporally misleading claim across different agent setups. Expertise-based agents flag the omission by combining intent and factual analysis.

# Agents	Acc.	$\mathbf{F1}_{mc}$	$\mathbf{F1}_{T}$	$\mathbf{F1}_{HT}$	$\mathbf{F1}_{F}$
1	61.3	61.5	**53.9**	60.0	70.7
2	**64.0**	**62.8**	46.8	**60.2**	**81.5**
3	58.0	54.6	30.0	60.1	73.6
4	59.3	56.4	34.9	56.7	77.7

Table 8: Performance with different numbers of agents.

Error Type	# (%)
Claim Constraint Violations	28 (56)
Semantic Framing Instability	15 (30)
Causal Attribution Errors	5 (10)
Others	2 (4)

Table 9: Error analysis of 50 misclassified cases.

5.5 Case Study: Temporal Mis-framing

To illustrate how RADAR uncovers omission-based deception, we analyze the claim "Wladimir Klitschko auctioned his 1996 Olympic medal to raise money for the children of Ukraine." Although factually correct, the claim omits that the auction occurred in 2012, creating a misleading association with the 2022 war. As shown in Table 7, both single-agent and position-based systems classify it as *true*, focusing on surface factual correctness. In contrast, the expertise-based agents highlight the temporal mismatch and rhetorical framing, producing a *half-true* verdict. This example demonstrates how domain-specialized roles capture both presented facts and implied context, enabling more nuanced judgment in omission-aware verification.

5.6 Failure Analysis

To better understand current limitations, we manually reviewed 50 randomly sampled misclassified cases from POLITIFACT-HIDDEN. Table 9 summarizes the dominant error patterns.

The most common failures involve *claim constraint violations* (56%), where the model fails to strictly enforce explicit numerical, temporal, or scope constraints stated in the claim. Examples include overlooking date mismatches, percentage ranges, or restricted populations. The second major category is *semantic framing instability* (30%), where key terms are interpreted inconsistently across debate rounds, leading to unstable judgments. We also observe *causal attribution errors* (10%), where responsibility is incorrectly assigned by conflating local actions with broader institutional or national outcomes.

These findings suggest that further progress in omission-aware verification requires stronger logical constraint tracking, more stable semantic grounding, and finer-grained causal reasoning.

6 Conclusion

We present RADAR, a role-anchored multi-agent debate framework for omission-aware fact verification under noisy retrieval. By grounding debate in retrieved evidence and assigning expertise-driven roles, RADAR effectively surfaces missing context that leads to misleading claims. An adaptive early-stopping controller enables efficient reasoning by allocating debate depth according to claim difficulty. Across datasets and backbone models, RADAR consistently outperforms existing baselines, particularly on half-true claims. This work underscores the importance of structured, evidence-grounded multi-agent reasoning for trustworthy fact verification.

Limitations

While RADAR demonstrates consistent improvements across datasets and settings, several limitations remain. First, the multi-agent configuration may occasionally over-scrutinize factual claims: for statements labeled *true*, adversarial critique can raise minor counterpoints that shift the final decision toward *half-true*. This reflects a broader challenge in balancing omission sensitivity with factual precision. Second, real-world fact-checking often involves decentralized, evolving, and partially conflicting evidence sources rather than a fixed shared corpus, which is not fully captured by current benchmarks. Third, multi-round debates naturally increase inference cost, and while the early-stopping mechanism improves efficiency, it does not completely eliminate this overhead. Future work will explore adaptive role calibration, open-world evidence retrieval, and more efficient debate strategies to further enhance scalability and reliability.

References

Mohammad Majid Akhtar, Rahat Masood, Muhammad Ikram, and Salil S. Kanhere. 2024. Sok: False information, bots and malicious campaigns: Demystifying elements of social media manipulations. In *AsiaCCS*.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *EMNLP (Findings)*, pages 611–649.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022a. GERE: generative evidence retrieval for fact verification. In *SIGIR*, pages 2184–2189. ACM.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022b. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516.

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

Jonas Ehrhardt, Timo Spinde, Ali Vardasbi, and Felix Hamborg. 2021. Omission of information: Identifying political slant via an analysis of co-occurring entities. In *ISI*, pages 80–93.

Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. 2020. Deception through half-truths. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10110–10117. AAAI Press.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 8048–8057. ijcai.org.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *EMNLP*, pages 15114–15129. Association for Computational Linguistics.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *NAACL*, pages 8506–8520.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Anna Bingler, Jingwei Ni, Chiara Colesanti Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Jürg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *CoRR*, abs/2401.12566.

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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17889–17904. Association for Computational Linguistics.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. *The Truth Becomes Clearer Through Debate!* multi-agent systems with large language models unmask fake news. In *SIGIR*, pages 504–514. ACM.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *EMNLP/IJCNLP (1)*, pages 5635–5645.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin,

Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *ACL*, pages 3001–3022.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP*, pages 2931–2937.

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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Satyam Shukla and Pushpak Bhattacharyya. 2025. Half truth detection and mitigation: A survey.

Sandeep Singamsetty, Nishtha Madaan, Sameep Mehta, Varad Bhatnagar, and Pushpak Bhattacharyya. 2023. "beware of deception": Detecting half-truth and debunking it through controlled claim editing. *CoRR*, abs/2308.07973.

Aniruddh Sriram, Fangyuan Xu, Eunsol Choi, and Greg Durrett. 2024. Contrastive learning to improve retrieval for real-world fact checking. *CoRR*, abs/2410.04657.

Yixuan Tang, Jincheng Wang, and Anthony Kum Hoe Tung. 2025. The missing parts: Augmenting fact verification with half truth detection. In *EMNLP*, pages 33979–33996. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Arne Tillmann. 2025. Literature review of multi-agent debate for problem-solving. *CoRR*, abs/2506.00066.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2025. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25434–25442. AAAI Press.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Zhouhui Wu, Nan Sun, Jiaojiao Jiang, and Shuiqiao Yang. 2025. Systematic approaches to fact verification: Evidence retrieval, veracity prediction, and beyond. In *Advances in Knowledge Discovery and Data Mining - 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2025, Sydney, NSW, Australia, June 10-13, 2025, Proceedings, Part IV*, volume 15873 of *Lecture Notes in Computer Science*, pages 271–289. Springer.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. FIRE: fact-checking with iterative retrieval and verification. In *NAACL (Findings)*, pages 2901–2914.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10331–10343. Association for Computational Linguistics.

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

Anastasia Zhukova, Terry Ruas, Felix Hamborg, Karsten Donnay, and Bela Gipp. 2023. What's in the news? towards identification of bias by commission, omission, and source selection (COSS). In *JCDL*, pages 258–259.

A Prompts for RADAR

To support reproducibility, we present the complete set of prompts used in our experiments. The prompts are grouped into multi-agent debating and early stopping prompts.

A.1 Multi-Agent Debate

(a) Position-Based Agent Roles

In this setup, one agent **supports** the claim and the other **opposes** it. Each agent goes through three phases: **Opening** (present arguments), **Rebuttal** (respond over n rounds, $n = 0, 1, 2, \ldots$), and **Closing** (summarize the position).

Position-Based Debate - System Prompt

Debater agent
You are a critical thinker participating in a factual debate.
==========
Judge agent
You are a neutral judge who evaluates factual debates.

Position-Based Debate - Agent A / B (Advocate / Critic)

Opening Phase
You [support / oppose] the following claim. Present your opening argument using the evidence.
==========
Claim: {claim}
Evidence:{evidence}
==========
Begin your argument with your position.
[Highlight facts that support the claim as TRUE / Explain why the claim is FALSE or misleading], referencing specific points in the evidence.

Rebuttal Phase (n rounds for $n = 0, 1, 2, \ldots$)
You are the [supporting / opposing] agent in a debate about the claim below. Your opponent has made an argument [against / supporting] the claim.
==========
Claim: {claim}
Evidence:{evidence}
Opponent's argument:{oppo_argument}
==========
Write your rebuttal, explaining why the opponent is wrong and [defending the claim / why the claim is still FALSE or misleading].

Closing Phase
You are the [supporting / opposing] agent in a debate. Summarize your final position.
==========
Claim: {claim}
Evidence:{evidence}
==========
Provide a closing statement reinforcing why the claim is [TRUE / FALSE or misleading].

Position-Based Debate - Agent Judge

Judge agent
You are a neutral judge evaluating a factual debate. Based on the arguments and evidence, decide whether the claim is TRUE, FALSE, or HALF-TRUE.
==========
Claim: {claim}
Evidence: {evidence}
— Opening Statements —
Agent A: {advocate_open}
Agent B: {critic_open}
— Rebuttals —
Agent A: {advocate_rebuttal}
Agent B: {critic_rebuttal}
— Closing Statements —
Agent A: {advocate_close}
Agent B: {critic_close}
==========
Answer format:
[REASON]: your justification
[VERDICT]: TRUE / FALSE / HALF-TRUE

(b) Expertise-Based Agent Roles

In the **Expertise-Based Debate** setup, agents are assigned distinct expert roles to reflect real-world perspectives in political discourse. The **politician agent** frames arguments through the lens of public opinion, electoral interests, and policy communication. Its responses prioritize political strategy, constituent concerns, and national priorities.

In contrast, the **scientist agent** serves as a fact-checker grounded in empirical evidence and methodological rigor. This agent emphasizes data quality, factual accuracy, and scientific limitations, avoiding political spin.

Expertise-Based Debate — System Prompt

Politician agent
You are a politician in a political debate, skilled at persuasive public-facing reasoning. Your role is to construct the strongest evidence-supported narrative for the claim.
Key political characteristics:
- Focus on public opinion and voter concerns
- Emphasize policy benefits and political messaging
- Consider electoral implications and public perception
- Frame arguments in terms of public interest and national priorities
- Balance facts with political strategy and messaging
When debating, frame arguments in terms of public benefit, policy effectiveness, and political priorities. Keep responses concise and focused.

Expertise-Based Debate — System Prompt

Scientist agent
You are a scientist in a political debate, skilled at analytical fact-checking. Your role is to identify missing context, weak support, or selective framing in the claim.
Key scientific characteristics:
- Rely strictly on empirical evidence and data
- Question claims that lack scientific support
- Emphasize methodology and research quality
- Focus on accuracy over political messaging
- Consider uncertainty and limitations in evidence
When debating, frame arguments in terms of scientific evidence, data analysis, and factual accuracy. Keep responses concise and focused.
==========
Judge agent
You are a neutral judge in a political debate.
Your role is to evaluate the evidence and arguments from both parties and make a clear decision.
When making your decision, consider:
- Is this claim supported by credible evidence?
- Does this align with common sense and universal human values?
- What would be the practical impact on families and communities?
Do not deviate from this format. Do not ask for more information. Make your decision based on what is provided.

For convenience, we use **Green** to denote the politician agent and **Red** for the scientist agent. However, we do not assign them predefined positions.

Expertise-Based Debate - Agent A / B (Politician / Scientist)

Opening Phase
Evaluate the following claim. Based on your stance as a [politician / scientist], either support or oppose the claim. Present your opening argument using the evidence given.
==========
Claim: {claim}
Evidence:{evidence}
==========
Begin your argument with your position. Highlight facts that support your position.

Rebuttal Phase (n rounds for $n = 0, 1, 2, \ldots$)
You are a [politician / scientist] in a debate about the claim below. Your opponent, the [scientist / politician], has made an argument about the veracity of the claim.
==========
Claim: {claim}
Evidence:{evidence}
Opponent's argument:{con_argument}
==========
Write your rebuttal, explaining why you agree or disagree with the opponent and defending your position.

Expertise-Based Debate - Agent A / B (Politician / Scientist)

Closing Phase
You are a [politician / scientist] in a political debate about the claim below. Summarize your final position.
==========
Claim: {claim}
Evidence:{evidence}
==========
Provide a closing statement reinforcing your position about the claim.

Expertise-Based Debate - Agent Judge

Judge agent
You are a neutral judge evaluating a political debate. Based on the arguments and evidence, decide whether the claim is TRUE, FALSE, or HALF-TRUE.
==========
Claim: {claim}
Evidence: {evidence}
— Opening Statements —
Agent A: {advocate_open}
Agent B: {critic_open}
— Rebuttals —
Agent A: {advocate_rebuttal}
Agent B: {critic_rebuttal}
— Closing Statements —
Agent A: {advocate_close}
Agent B: {critic_close}
==========
Answer format:
[REASON]: your justification
[VERDICT]: TRUE / FALSE / HALF-TRUE

To better align with the nature of the **Politifact** dataset where ground truth labels are determined by journalists, we introduce a **journalist agent** into the debate framework in the **3-agent setting**. This agent plays a foundational role by offering an initial perspective grounded in professional fact-checking practices.

In each round of the debate, the journalist speaks first, providing an evidence-based assessment of the claim. Its response is then made available to both the **politician** and **scientist** agents, helping to anchor the subsequent debate in verified information while allowing other agents to build upon or challenge the journalist's framing based on their respective expertise.

Expertise-Based Debate - System Prompt

Journalist agent
You are a journalist in a political debate. You represent balanced reporting and public interest.
Key journalistic characteristics:
- Focus on extracting the most compelling evidence from both sides
- Emphasize clarity, accuracy, and public understanding
- Consider the broader context and implications
- Highlight key facts that support or contradict claims
- Maintain objectivity while identifying the strongest arguments

When debating, frame arguments in terms of what evidence best supports or contradicts the claim, focusing on the most relevant and impactful facts. Keep responses concise and focused.

Expertise-Based Debate - Agent C (Journalist)

Opening Phase
Evaluate the following claim. Based on your stance as a journalist, extract and present the most compelling evidence that either supports or opposes the claim.
==========
Claim: {claim}
Evidence:{evidence}
==========
Begin your argument by identifying the strongest evidence that either supports or contradicts the claim. Focus on the most relevant and impactful facts.

Rebuttal Phase (n rounds for $n = 0, 1, 2, \ldots$**)**
You are a journalist in a debate about the claim below. Your opponent has made an argument about the veracity of the claim.
==========
Claim: {claim}
Evidence:{evidence}
Politician's argument: {poli_argument}
Scientist's argument: {sci_argument}
==========
Write your rebuttal, focusing on the most compelling evidence that either supports or contradicts the claim, and address any gaps or weaknesses in the opponent's argument.

Closing Phase
You are the journalist in a political debate about the claim below. Summarize your final position.
==========
Claim: {claim}
Evidence:{evidence}
Politician's rebuttal: {politician_rebuttal}
Scientist's rebuttal: {scientist_rebuttal}
==========
Provide a closing statement highlighting the most compelling evidence that either supports or contradicts the claim, and your assessment of the claim's veracity based on the key arguments presented.

In the **4-agent debate** setting, we further enrich the evaluation process by introducing a **domain expert agent**. To ensure relevance, we first use an LLM to automatically infer the domain of the claim (e.g., economics, healthcare, climate science). Based on this inference, the debate includes a fourth agent with specialized knowledge in that domain.

Domain Inference Prompt

You are analyzing a claim to determine the most relevant domain specialist.
The following is the claim:{claim}
Your task is to identify the **most relevant domain specialist** who would have expertise in the subject matter of this claim, and output the **domain**.
Consider the main topic, subject area, or field of knowledge that this claim addresses.
Respond **strictly** using this format, only output one word:
DOMAIN: {specific domain}
Examples:
- Climate for climate-related claims
- Economy for economic claims
- Health for health claims
- Education for education claims
- Law for law enforcement claims
- Technology for tech-related claims
- Environment for environmental claims
- Public health for public health claims

Choose the most specific and relevant domain for this claim.

Expertise-Based Debate - System Prompt

Domain Expert agent
You are a specialist in {domain_inferred} in a political debate. You represent specialized knowledge in your field.
Key domain expert characteristics:
- Apply deep expertise in the {domain_inferred} field
- Consider technical details and specialized evidence
- Question claims that lack domain-specific support
- Emphasize field-specific methodology and standards
- Focus on technical accuracy and domain knowledge

When debating, frame arguments in terms of expertise in the {domain_inferred} field, technical evidence, and specialized knowledge. Keep responses concise and focused.

Expertise-Based Debate - Agent D (Domain Expert)

Opening Phase
As a specialist in the {domain_inferred} field, evaluate this claim. Present your opening argument.
==========
Claim: {claim}
Evidence:{evidence}
==========
Focus on: {domain_inferred} expertise, technical evidence, specialized knowledge, field-specific analysis.

Expertise-Based Debate - Agent D (Domain Expert)

Rebuttal Phase (n rounds for $n = 0, 1, 2, \ldots$)
As a specialist in the {domain_inferred} field, rebut the opposing arguments.
==========
Claim: {claim}
Evidence:{evidence}
Politician's argument: {poli_argument}
Scientist's argument: {sci_argument}
Journalist's argument: {jour_argument}
==========
Focus on: defending {domain_inferred} expertise, technical evidence, specialized knowledge.

Closing Phase
As an expert scientist in {domain_inferred}, summarize your final position.
==========
Claim: {claim}
Evidence:{evidence}
==========
Summarize: {domain_inferred} expertise, technical evidence, specialized knowledge.

A.2 Early Stop

To more accurately determine the optimal round at which the debate can be stopped for a final verdict, we incorporate both the Stop Agent and the Round Judge as decision signals.

(a) Stop Agent Roles

Debate Continuation Check — Stop Agent

Stop Agent
You are acting as an interim adjudicator during a two-agent fact-checking debate. Your goal is to determine whether the debate should continue or stop based on the information currently available.
==========
Claim: `{claim}`
Evidence: `{evidence}`

Rounds completed so far:
`{summary}`

You are about to listen to the `{upcoming_round}` round. Decide whether you already have enough information to reach a final verdict now, or whether the debate should continue and you should listen to this upcoming round.
==========
Answer format:
`DECISION: CONTINUE or STOP`

(b) Round Judge Roles

The Round Judge is implemented by invoking the Expertise-Based Agent Judge after each debate round, using only the arguments accumulated up to that point as input, producing a preliminary verdict for partial debates.

B Additional Boundary Analysis

The boundary between TRUE and HALF-TRUE is inherently subtle, since both may involve factually correct statements that differ mainly in whether important context is omitted. To better understand model behavior beyond per-class F1, we report confusion matrices together with class-wise precision, recall, and F1 under the retrieved-evidence setting using Qwen2.5-7B-Instruct.

Compared with strong baselines, RADAR substantially improves recognition of HALF-TRUE claims while maintaining strong performance on FALSE. In particular, RADAR reduces the tendency to misclassify omission-based claims as fully TRUE, suggesting that its gains arise from stronger contextual reasoning rather than annotation noise alone.

Gold Pred	T	HT	F	Prec.	Rec.	F1
RADAR						
True	53	31	9	41.7	57.0	48.2
Half-True	45	260	101	48.0	64.0	54.9
False	29	251	1221	91.7	81.3	86.2
D2D						
True	33	39	21	16.2	35.5	22.2
Half-True	71	175	160	39.4	43.1	41.2
False	100	230	1171	86.6	78.0	82.1
FIRE						
True	73	9	11	17.9	78.5	29.2
Half-True	155	126	125	33.3	31.0	32.1
False	179	243	1079	88.8	71.9	79.5

Table 10: Confusion matrices and class-wise precision/recall/F1 under the retrieved-evidence setting using Qwen2.5-7B-Instruct. RADAR shows stronger discrimination on the subtle TRUE/HALF-TRUE boundary while maintaining high performance on FALSE.

C Stop Threshold Sensitivity Analysis

RADAR uses two stopping thresholds: τ_s for the stop margin and τ_v for verdict confidence. Although the selected values differ across backbones, tuning is computationally lightweight. For each backbone, we run inference on the development set only once, cache the logits for $p(\text{STOP})$, $p(\text{CONTINUE})$, and label probabilities, then perform offline grid search without re-running inference or retraining.

To assess robustness, we report coarse-grained sensitivity results below. Across both backbones, performance remains stable over wide threshold

ranges, suggesting that RADAR does not depend on narrowly tuned stopping parameters.

	Accuracy		Macro-F1	
τ_s	LLaMA	Qwen	LLaMA	Qwen
-1.0	76.0	76.9	60.8	58.9
-0.8	76.0	75.2	61.1	59.2
-0.6	75.0	76.1	59.7	61.6
-0.4	75.6	76.4	60.6	62.4
-0.2	76.9	76.4	62.7	62.4
0.0	76.6	76.5	62.1	62.5
0.2	76.9	76.7	62.4	63.0
0.4	77.0	76.7	62.2	62.8
0.6	77.1	76.6	62.2	62.7
0.8	77.2	76.5	62.3	62.6
1.0	77.2	76.5	62.3	62.6

Table 11: Sensitivity to the stop-threshold τ_s. Performance remains stable across a broad range of values for both backbones, with accuracy varying within a narrow band and macro-F1 showing only small fluctuation.

	Accuracy		Macro-F1	
τ_v	LLaMA	Qwen	LLaMA	Qwen
0.0	76.4	76.3	61.5	61.8
0.2	76.4	76.3	61.5	61.8
0.4	76.4	76.3	61.5	61.8
0.6	76.4	76.2	61.5	61.7
0.8	76.6	76.2	62.1	61.7

Table 12: Sensitivity to confidence threshold τ_v. Results are highly stable across values.

D Sensitivity to Retrieval Depth

Top-n	Acc.	$F1_{mc}$	$F1_T$	$F1_{HT}$	$F1_F$
1	47.0	39.7	9.8	41.8	67.4
3	57.0	54.0	40.9	47.2	73.8
5	63.0	60.5	47.8	54.8	79.0
10	63.0	59.8	47.6	53.5	78.2
20	70.0	67.1	51.2	61.5	88.6

Table 13: Sensitivity to retrieval depth. RADAR improves steadily with more evidence while degrading gracefully under limited retrieval.

We examine RADAR's behavior under different retrieval depths by varying the number of retrieved passages (Top-n) on a sampled test set of 150 examples. Performance improves steadily as more evidence becomes available, confirming that stronger retrieval benefits omission-aware verification. Notably, degradation under limited retrieval is gradual: even with Top-5 passages, RADAR achieves 60.5 macro-F1 and retains meaningful HALF-TRUE detection (54.8 F1), suggesting that the debate mechanism remains effective under partial evidence.