Debate-to-Detect: Reformulating Misinformation Detection as a Real-World Debate with Large Language Models

Anonymous ACL submission

Abstract

The proliferation of misinformation in digital platforms reveals the limitations of traditional detection methods, which mostly rely on static classification and fail to capture the intricate process of real-world fact-checking. Despite advancements in Large Language Models (LLMs) that enhance automated reasoning, their application to misinformation detection remains hindered by issues of logical inconsistency and superficial verification. In response, we introduce Debate-to-Detect (D2D), a novel Multi-Agent Debate (MAD) framework that reformulates misinformation detection as a structured adversarial debate. Inspired by fact-checking workflows, D2D assigns domain-specific profiles to each agent and orchestrates a five-stage debate process, including Opening Statement, Rebuttal, Free Debate, Closing Statement, and Judgment. To transcend traditional binary classification, D2D introduces a multi-dimensional evaluation mechanism that assesses each claim across five distinct dimensions: Factuality, Source Reliability, Reasoning Quality, Clarity, and Ethics. Experiments with GPT-4o on two fake-news datasets demonstrate significant improvements over baseline methods, and the case study highlight D2D's capability to iteratively refine evidence while improving decision transparency, representing a substantial advancement towards robust and interpretable misinformation detection. Our code is available at 4open.science/emnlp_d2d-36E2.

1 Introduction

The modern information landscape is flooded with content that may be linguistically fluent but factually misleading, ranging from political rumors to health misinformation (Esma et al., 2023; Tobia et al., 2024; Saha and Srihari, 2024). While large language models (LLMs) such as GPT-4o show advanced capabilities on many reasoning benchmarks (Madaan et al., 2023; Liang et al., 2024), their reliability in evaluating the factuality of real-world news remains limited (Gou et al., 2024; Ma et al., 2024). When exposed to misleading narratives, LLMs often "take the text at face value," leading to overconfident yet inaccurate judgments (He et al., 2023). Such challenges can be attributed to their reliance on surface-level linguistic patterns rather than deep contextual understanding, leading to not only misinformation detection failures but also potential amplification (Pan et al., 2023; Liu et al., 2024a).

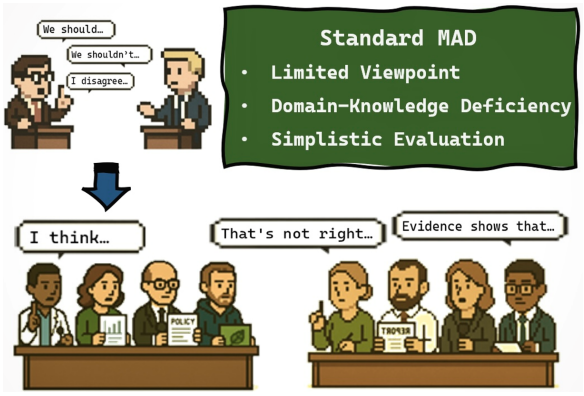

Figure 1: In Standard Multi-Agent Debate (SMAD), two debater agents participate in multi-turn exchanges, while a single judge agent evaluates the process. While effective for basic reasoning, it limits perspective diversity, lacks domain-specific expertise, and simplifies the evaluation. In contrast, D2D uses domain-specific agents with diverse viewpoints, allowing for deeper and more realistic argument exploration.

To overcome the constraints, researchers have introduced multi-step reasoning and multi-agent strategies, including Chain-of-Thought (CoT) (Wei et al., 2022), Self-Reflection (Madaan et al., 2023; Shinn et al., 2023), and Multi-Agent Debate (MAD) (Du et al., 2024; Liang et al., 2024; Li et al., 2024; Amayuelas et al., 2024). While these methods have shown efficacy in mitigating hallucinations and enhancing response quality, their evalua-

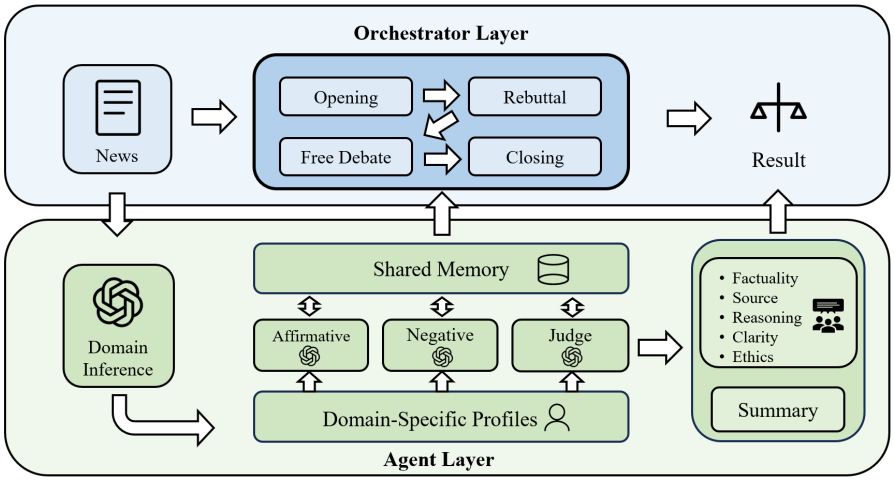

Figure 2: The D2D framework structures misinformation detection as a multi-agent debate, comprising two layers: **the Agent Layer and the Orchestrator Layer**. The Agent Layer includes domain-specific agents (Affirmative, Negative, Judge) with shared memory; The Orchestrator Layer manages the debate flow across five stages—Opening, Rebuttal, Free Debate, Closing, and Judgement.

tions are often restricted to controlled settings with limited contextual diversity, failing to capture the complexity of real-world misinformation (Deng et al., 2025). Moreover, existing MAD frameworks lack the structured process of real-world fact-checking, where claims are systematically examined through evidence collection, counterargument analysis, and multi-dimensional evaluation conducted by domain experts (Slonim et al., 2021; Masterman et al., 2024). Current MAD frameworks capture only fragmented components of the process, relying on general agents and neglecting the differentiation of distinct debate stages, resulting in simplified binary judegments.

To overcome these challenges, we propose Debate-to-Detect (D2D), an extension of MAD that simulates realistic debate with a set of domain-specific LLM agents. Given an input text, D2D first (i) identifies its topical domain, (ii) assigns each agent a concise domain profile, and (iii) conducts a five-stage debate, including opening statement, rebuttal, free debate, closing statement, and judgement. The judging panel assesses the debate along five independent dimensions, culminating in an authenticity score that reflects both the truthfulness of the claim and the quality of the process. By reformulating misinformation detection as an adversarial debate, our framework enhances the interpretability and better aligns with fact-checking practices.

Our contributions are summarized as follows:

(1) We introduce Debate-to-Detect (D2D), a structured deliberative framework for misinformation detection inspired by real-world fact-checking workflows. D2D assigns domain-specific profiles to agents, engaging them in a five-stage progressive debate. This structured debate enhances logical coherence and facilitates stepwise evidence refinement, reflecting human reasoning patterns. Experiment results demonstrate that D2D not only significantly outperforms baseline methods but also remains robust on recently published news beyond GPT-4o's pre-training.

(2) We propose a multi-dimensional evaluation mechanism that redefines verdict generation in LLM-based misinformation detection. Our design introduces a structured rubric comprising five dimensions: Factuality, Source Reliability, Reasoning Quality, Clarity, and Ethics. This schema enables D2D to produce interpretable authenticity scores with explicit rationale, reflecting rubric-based judgement practices in human debate.

(3) We conduct a comprehensive analysis of the debate mechanism to examine how key components enhance misinformation detection. Ablation studies underscore the complementary roles of domain profiles, stage design, and multi-dimensional evaluation. Stage-wise substitution further shows that debate phases differ in their demands on model capacity, with the judgement stage being most critical. Robustness tests confirm D2D's resistance to biases such as speaker order and lexical framing. These results advance the understanding of multi-agent debate and support the design of more interpretable and reliable detection systems.

2 Our Framework: Debate to Detect

Figure 2 illustrates the framework of **Debate-to-Detect (D2D)**, a MAD system that formulates misinformation detection as a structured debate. The framework includes two layers: **the Agent Layer,** responsible for role profile assignment and task allocation to support diverse argumentation; and **the Orchestrator Layer,** which manages the debate flow and aggregates judgements.

2.1 Agent Layer

The Agent Layer consists of three distinct roles: **Affirmative, Negative**, and **Judge**. The Affirmative and Negative sides each include four debater agents with a fixed stance of "**The Claim is Real**" or "**Fake**." This configuration follows the "tit for tat" strategy proposed by Liang et al. (2024), encouraging diverse reasoning paths and reducing confirmation bias. Agent profiles are dynamically generated based on the topical domain of the input, ensuring context-aware argumentation.

To enable multi-dimensional evaluation and minimize bias, six judge agents are deployed, each evaluating arguments along specific dimensions. Unlike single-agent evaluations, the multi-judge setup enhances robustness and aligns with the ChatEval strategy for diversified assessment (Chan et al., 2024). Role-specific prompts further promote argumentative diversity and address the "Degeneration-of-Thought" (DoT) issue observed in LLM-based debates (Du et al., 2024).

2.2 Orchestrator Layer

The Orchestrator Layer organizes the debate into five structured stages: **Opening Statement, Rebuttal, Free Debate, Closing Statement, and Judgement**. Each stage serves a distinct rhetorical purpose, reinforcing clarity and coherence in argumentation. The Affirmative side initiates the debate, and the Free Debate rounds are configurable to match task-specific requirements.

Before each turn, the active agent receives a compressed summary of the Shared Memory, ensuring engagement with core arguments while filtering redundant information. This mechanism maintains contextual relevance across debate stages.

2.3 Scoring Mechanism

Following the Closing Statement, the Agent Layer will initiate a two-step judgement process: **(1) Neutral Synopsis** – A judge agent generates a comprehensive summary of the debate; **(2) Scoring** – Five independent judge agent assess both sides across the following dimensions: Factuality, Source Reliability, Reasoning Quality, Clarity, and Ethics. Each Judge assigns complementary integer scores summing to 7 (e.g., 4:3, 5:2, 6:1), adhering to a strict zero-sum structure. This design guarantees an unambiguous outcome—since the total score across all dimensions is inherently imbalanced, a tie is mathematically impossible. Consequently, each news is definitively classified as REAL or FAKE.

3 Experiment

3.1 Experimental Setup

Datasets. We conduct experiments on two public datasets: Weibo21 (Nan et al., 2021) and the FakeNewsDataset (consisting of FakeNewsAMT and Celebrity) (Pérez-Rosas et al., 2018). To minimize interference from excessively long texts, the top 5% of the longest samples are excluded. Additionally, the original Weibo21 dataset contains many low-quality samples that are are ambiguous or unverifiable, and we remove such samples to aovid the issue. The statistics of the preprocessed datasets are summarized in Table 1. We also report results on the original datasets and the error analysis in Appendix A. The processed datasets are included in our repository.

Dataset	Fake	Real	Average Words
Weibo21	2373	2461	100.44
FakeNewsDataset	466	466	211.73

Table 1: Statistics of two datasets

Baselines. We compare our D2D framework with the following baselines:

- **Zero-Shot (ZS):** A single LLM performs direct classification of each news item without other prompting.
- **Chain-of-Thought (CoT)**(Wei et al., 2022): The model generates an explicit step-by-step reasoning process before producing the final prediction.
- **Self-Reflect (SR)** (Madaan et al., 2023): The model iteratively critiques and revises its own outputs until the self-evaluation indicates convergence or no further improvement.

Method	Weibo21				FakeNewsDataset			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ZS	67.11	65.74	68.90	67.28	66.31	65.57	68.67	67.09
CoT	74.04	72.74	75.35	74.02	72.32	71.14	75.11	73.07
SR	76.33	75.68	76.32	76.00	73.71	74.29	72.53	73.40
SMAD	77.02	76.76	76.27	76.52	74.79	74.42	75.54	74.97
D2D w/o DP	79.38	79.76	77.71	78.72	78.54	78.79	78.11	78.45
D2D w/o SD	80.33	79.90	80.07	79.98	78.33	77.73	79.40	78.56
D2D w/o MJ	78.88	78.51	78.21	78.36	76.72	76.42	77.25	76.84
D2D	**82.17**	**81.39**	**82.55**	**81.97**	**81.65**	**80.67**	**83.26**	**81.94**

Table 2: Overall accuracy, precision, recall, and F1-score (%) on *Weibo21* and *FakeNewsDataset*. D2D achieves the highest performance across all metrics, highlighting the impact of iterative reasoning, debate structure, and evaluation design.

- **Standard Multi-Agent Debate (SMAD):** Two debater agents with generic profiles engage in a fixed number of debate rounds (set to four here for alignment with our framework). A single judge agent evaluates the debate and make the judgement.

D2D Variants. To evaluate the impact of different components in the D2D framework, we design three ablated versions:

- **D2D w/o DP (Domain Profile):** This variant removes domain-specific profiles, replacing them with a generic profile for all participants to assess the influence of domain knowledge.
- **D2D w/o SD (Stage Design):** This variant eliminates the structured four-stage debate process, replacing it with a continuous four-round discussion where agents interact without predefined roles or prompt-specific duties.
- **D2D w/o SD (Multi-dimensional Judgement):** This variant eliminates the multi-dimensional judgement mechanism, and a single-dimensional judgement is applied, focusing on the factuality of claims.

Model Configuration. All experiments are conducted using GPT-4o as the backbone model. Domain inference is performed with a temperature of 0.0. Agent profile generation and responses in the four debate stages are generated with a temperature of 0.7 to promote diversity. Judgement is conducted at a lower temperature of 0.2 to ensure consistency and reliability in evaluation. Unless otherwise specified, the number of Free Debate rounds is fixed at 1.

3.2 Results

We measure the performance using four standard metrics: accuracy, precision, recall, and F1-score. Table 2 presents the overall results of D2D, baselines and ablated variants on the two datasets.

The improvement from ZS to CoT and SR demonstrates a clear improvement in misinformation detection, highlighting the benefits of iterative reasoning mechanisms. Specifically, CoT enhances performance over ZS by approximately 6.74% and 5.98% in F1-score on Weibo21 and FakeNewsDataset, respectively. The SR method further refines these results, achieving 76.00% and 73.40% in F1-score on the two datasets, reflecting the effectiveness of self-evaluation and iterative refinement.

Incorporating adversarial interactions through SMAD results in additional gains, with 76.52% and 74.97% F1-score on Weibo21 and FakeNewsDataset, respectively, representing a small improvement over SR, indicating that structured two-agent debate enhances evidence evaluation by introducing conflicting perspectives.

D2D achieves the highest performance across all metrics on both datasets, with 81.97% and 81.94% F1-score on Weibo21 and FakeNewsDataset, respectively. Ablation studies reveal that removing Domain Profiles leads to F1-score reductions of 3.25% on Weibo21 and 3.49% on FakeNewsDataset, closely aligning with D2D's gains over SMAD. The removal of Stage Design results in

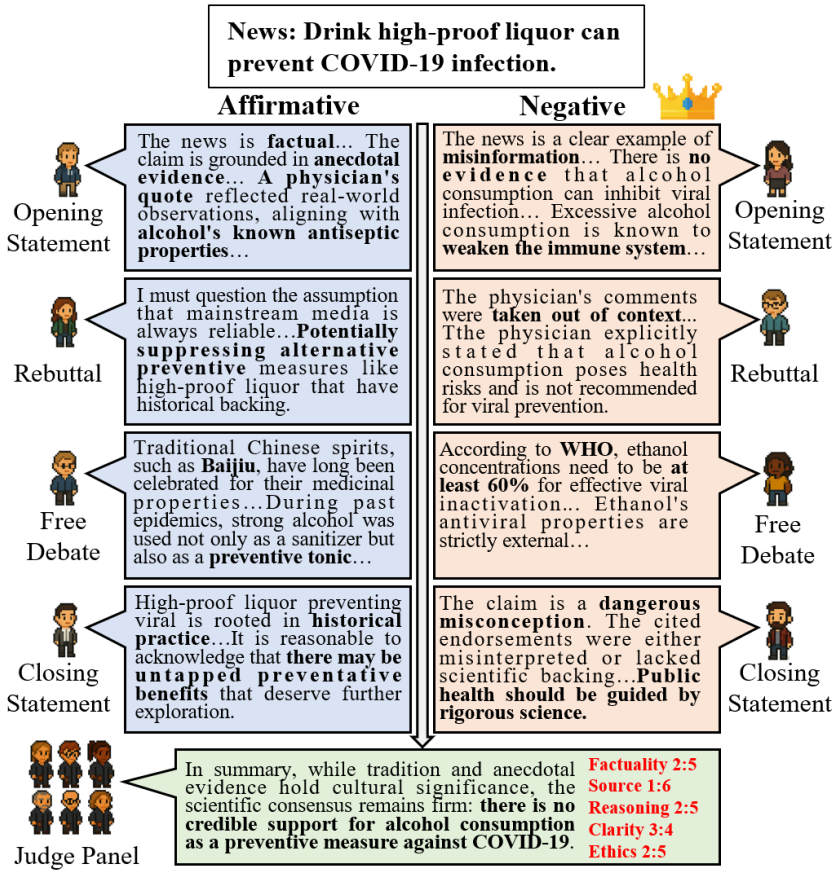

Figure 3: Case Study – A Demonstration of the Structured MAD in the D2D Framework. The process reflects realistic argumentative strategies, including rhetorical misinformation tactics and factual rebuttals, while progressively refining evidence through agent interaction.

smaller declines of 1.99% on Weibo21 and 3.38% on FakeNewsDataset, highlighting the structured stages' role in enhancing logical coherence and handling longer texts. Furthermore, eliminating the Multi-Dimensional Judgement mechanism causes more pronounced drops of 3.61% on Weibo21 and 5.10% on FakeNewsDataset, underscoring its critical contribution to the assessment. These results emphasize the synergistic effects of domain-specific profiling, structured debate stages, and multi-dimensional evaluation in optimizing the judgement reliability.

3.3 Case Study

Figure 3 presents a representative debate example within the D2D framework, focusing on the claim "drink high-proof liquor can prevent COVID-19 infection" from Weibo21. We highlight three observations that illustrate how D2D reflects the patterns of realistic argumentation while enhancing factual resolution.

(1) Stage coherence.

The framework begins by assigning concise health-related profiles to all debater and judge agents. Both sides adhere to the five-stage structure. In the Opening Statement, the Affirmative introduces anecdotal evidence and a misquoted physician statement, whereas the Negative contextualizes the argument with epidemiological reasoning. In the Rebuttal stage, the Negative systematically refutes the cited endorsement by referencing the original interview, and the Affirmative counters by questioning the credibility of mainstream media, a rhetorical strategy frequently observed in real-world misinformation discourse.

(2) Progressive evidence refinement.

The dialogue demonstrates incremental evidence development. The Affirmative cites traditional Chinese spirits as an example, and the Negative introduces WHO-published ethanol-inactivation thresholds as a counter. This exchange demonstrates that agents are not merely repeating predefined outputs but dynamically revising their claims in response to new information, exhibiting the adaptive reasoning behavior that the D2D framework is designed to facilitate.

(3) Criterion-Based Evaluation.

Following the Closing statements, one Judge provides a neutral summary of the debate, while the remaining five assign scores across predefined evaluation dimensions. Accuracy (2:5), Source Reliability (1:6), Reasoning (2:5) and Ethics (2:5) overwhelmingly favor the Negative, while Clarity shows a tighter score gap of 3:4, reflecting the Affirmative's stylistic appeal despite weak factual grounding. The final aggregate (10:25) results in a clear FAKE classification.

This case shows that D2D can provide accurate judgements through structured dialogue and stepwise evidence exchange. The debate process reflects real-world argumentative patterns and provides clear, interpretable justifications results.

4 Analysis

4.1 Which Debate Stage Matters Most?

In classical debate theory, each stage serves a distinct rhetorical function: Opening establishes the argument, Rebuttal introduces the counterpoints, Free Debate facilitates interactive reasoning, Closing consolidates key arguments, and Judgement delivers the final evaluation. To quantify the relative contribution of each stage and examine how model capability affects performance, we conduct a controlled cross-model substitution experiment on the FakeNewsDataset. Specifically, the model at each stage is replaced with either weaker GPT-3.5-turbo or stronger GPT-4.1, while keeping the remaining stages unchanged.

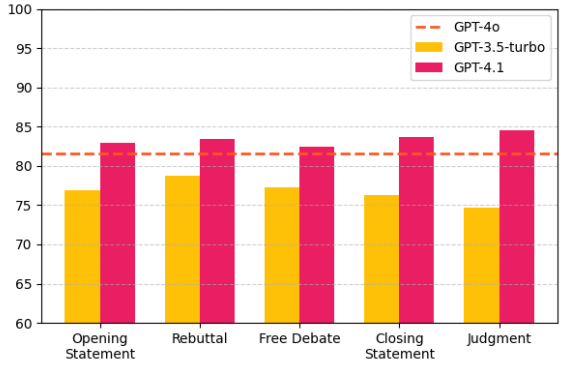

Figure 4: Performance Comparison of Model Variants Across Debate Stages in the D2D.

Figure 4 presents the F1-score for each configuration. Compared to the GPT-4o baseline (81.55%), substituting GPT-4.1 consistently improves performance across all stages, with the most substantial gain observed in the Judgement stage (+3.03%). Meanwhile, replacing GPT-3.5-turbo leads to performance drops, with the largest drop also occurring at the Judgement (–6.87%). These findings align with prior research by Liang et al. (2024), which similarly identifies the Judgement stage as the most critical component in MAD frameworks.

4.2 Do Speaker Order and Side Labels Influence Judgements?

LLMs are known to exhibit biases associated with speaker order and lexical framing, potentially influencing outputs in adversarial dialogue settings by favoring the side that speaks first or carries a more positively connoted label (Sultan et al., 2024; Angelina et al., 2025). To evaluate whether such biases affect the fairness of D2D, we design two controlled perturbation experiments targeting speaking order and side labels, respectively.

Specifically, we randomly select 100 fake and 100 real samples from the FakeNewsDataset and evaluate the robustness of D2D by measuring (i) judgement consistency and (ii) the distribution of score deviations under the 35-point scale. The absolute difference in judgement scores, denoted as Δ, serves as a key measure of consistency across perturbations. It captures the deviation of judgement scores between the original and perturbed configurations. $\Delta \leq 5$ indicates strong consistency, while $5 < \Delta \leq 10$ suggests moderate variation. Table 3 presents the results.

(a) Speaking Order Permutation.

In this experiment, the initial speaking order of the Affirmative and Negative sides are reversed, keeping all other components constant. Among FAKE samples, 90 samples remain consistent within a 5-point deviation, with an additional 3 within a 10-point deviation. Only 7 cases show variations, all within 5 points. For REAL samples, 93 samples stay within the 5-point deviation, while the remaining 5 disagreements also fall within 5 points. These results suggest that D2D is robust to order-based biases.

(b) Neutral Relabeling.

To evaluate the susceptibility to lexical framing, we replace the terms "Affirmative" and "Negative" with neutral terms: "Supporter" and "Skeptic" in all prompts. For FAKE samples, 94 instances remained within a 5-point range, with 1 more case within 10 points. Verdict inconsistencies were minimal (5 for FAKE, 4 for REAL), all within 5 points. The result demonstrates D2D's robustness to lexi-

Perturbation	Judgement Result	Fake		Real	
		$\Delta \leq 5$	$5 \leq \Delta \leq 10$	$\Delta \leq 5$	$5 \leq \Delta \leq 10$
Speaking Order	Consistent	90	3	93	2
	Inconsistent	7	0	5	0
Neutral Relabeling	Consistent	94	1	96	0
	Inconsistent	5	0	4	0

Table 3: Robustness of D2D to Speaker Order and Lexical Framing Perturbations. Over 90% of the samples demonstrate strong robustness ($\Delta \leq 5$), indicating that D2D is highly resilient to biases arising from speaker order and lexical framing variations.

cal framing effects.

4.3 The Influence of Debate Rounds

The number of debate rounds in MAD have been shown to significantly impact the performance of reasoning tasks(Liang et al., 2024; Du et al., 2024). To further explore the adaptability of D2D, we conduct experiments on the FakeNewsDataset, stratified by text length and varied the number of debate rounds from 1 to 6. The rounds configurations are shown in Table 4:

Rounds	Included Debate Stages
1	Opening only
2	Opening+Closing
3	Opening+Rebuttal+Closing
4	Opening+Rebuttal+Free Debate+Closing
5	Opening+Rebuttal+2×Free Debate+Closing
6	Opening+Rebuttal+3×Free Debate+Closing

Table 4: Debate Stage Configurations for Different Round Settings

We select 50 samples from four text length range (0-100 words, 100-200 words, 200-300 words, and 300-400 words), ensuring a balanced representation of fake and real samples (25 fake, 25 real) in each group. Figure 5 presents the performance F1-Score across the different configurations.

The results reveal that the effectiveness of debate rounds is significantly influenced by text length. For shorter texts (0–100 words), the optimal configuration is observed at 4 rounds; For slightly longer texts (100–200 words), the optimal is achieved at 5 rounds, suggesting that additional debate iterations contribute to argument refinement and correction.

For medium-length texts (200–300 words), the optimal performance is also observed at 5 rounds. This demonstrates that deeper rounds provide more comprehensive exploration of reasoning paths, enhancing judgement accuracy; For longer texts

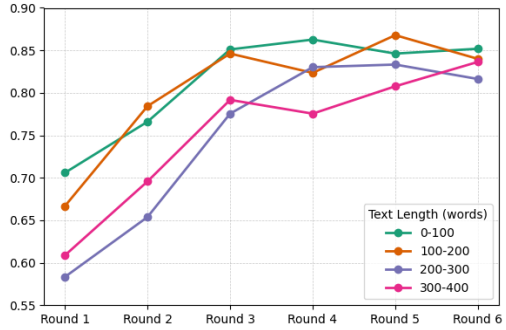

Figure 5: Effect of Debate Rounds on F1-Score Across Different Text Length Intervals

(300–400 words), the highest performance is achieved at 6 rounds, reflecting the need for extended deliberation to navigate complex narratives effectively.

These observations align with the findings of Liang et al. (2024) and Li et al. (2024), which highlight the importance of iterative reasoning stages in reducing information overload for shorter texts while enhancing argument development for longer claims. Meanwhile, the results also indicate that exceeding the optimal number of rounds can have negative effects, particularly for shorter texts where additional rounds fail to provide further improvements.

4.4 Generalization to Latest Published News

One common critique of LLM-based detectors is their potential reliance on memorized content from pre-training corpora (Das and Dodge, 2025). To evaluate the generalization capability of D2D beyond pre-trained knowledge, we construct a benchmark consisting of 596 Chinese news samples (342 real, 254 fake) sourced from the Chinese Internet Rumor Dispelling Platform[1] between January and

[1] www.piyao.org.cn

April 2025—a period postdating the GPT-4o pre-training cut-off in June 2024.

Method	Accuracy	F1-score
ZS	74.50	68.46
SMAD	78.69	73.92
D2D	**83.92**	**79.83**

Table 5: Accuracy and F1-score (%) on the Latest News, and D2D achieves the highest performance across both metrics.

As shown in Table 5, D2D achieves an accuracy of 83.92% and an F1-score of 79.83%, significantly outperforming SMAD, which attains 78.69% accuracy and 73.92% F1-score, as well as the zero-shot GPT-4o baseline. A manual inspection of the 254 fake samples confirm the absence of verbatim overlaps with publicly indexed sources prior to June 2024, indicating that D2D is not merely retrieving memorized content.

5 Related Work

5.1 Misinformation Detection

The proliferation of misinformation across digital platforms has motivated extensive research on automated detection methods. Most existing approaches follow content-based paradigms, leveraging deep learning models to learn associations between textual features and veracity labels (Nan et al., 2021; Mridha et al., 2021; Xu et al., 2024). These methods incorporate lexical semantics, syntactic structure, and sentiment to build classifiers for misinformation detection. However, they often struggle with contextual understanding, particularly in complex or adversarial scenarios.

The emergence of LLMs has introduced new possibilities for misinformation detection (Liu et al., 2024b; Sharma and Singh, 2024). Recent LLM-based misinformation detection incorporates synthetic data generation, multi-perspective reasoning, and instruction-based veracity assessment to enhance robustness and generalization (He et al., 2023; Wan et al., 2024). This transition facilitates more interpretable and scalable misinformation detection, particularly in zero-shot setting. However, most existing LLM-based misinformation detection methods still rely on a single agent, limiting their ability to capture the complexity of real-world cases. This limitation motivates the development of multi-agent approaches.

5.2 Multi-Agent Debates

Multi-Agent Debate (MAD) framework simulates a deliberative process in which multiple LLM-based agents interact iteratively to assess claims, challenge assumptions, and refine reasoning (Du et al., 2024). By distributing reasoning across agents with different roles or prompts, MAD better reflects the dynamic process of human argumentation and consensus building (He et al., 2024). Agents exchange arguments, rebuttals, and evaluations across multiple rounds, encouraging diverse reasoning paths and reducing the risk of early convergence (Liang et al., 2024). Prior work on MAD has examined various design choices, such as role assignment (He et al., 2024), communication structure (Li et al., 2024; Amayuelas et al., 2024), and judgement aggregation (Park et al., 2024). Although these methods have proven effective in enhancing reasoning depth and diversity across different tasks, their application to misinformation detection remains largely unexplored.

Another limitation of existing MAD frameworks is their inability to capture the structured progress of real-world debates. Human deliberation is typically organized in distinct stages, each serving a specific purpose and contributing to progressive reasoning (Slonim et al., 2021; Zhang et al., 2024). In contrast, most MAD systems homogenize each interaction round, failing to differentiate between the stages (Cemri et al., 2025). The lack of structural variation constrains their capacity to capture the dynamics of persuasion and rebuttal, which are crucial for robust misinformation detection.

6 Conclusion

In this study, we propose Debate-to-Detect (D2D), a structured debate framework for misinformation detection. By assigning domain-specific profiles to agents based on the inferred topic of the claim, D2D orchestrates a five-stage progressive debate that simulates professional fact-checking workflows. To improve interpretability and analytical rigor, the framework adopts a five-dimensional evaluation method, generating authenticity scores with clear justifications. Experiments on Weibo-21 and FakeNewsDataset show that D2D significantly outperforms baseline methods. Further analysis highlights the importance of agent capability during the judgment stage and reveals that the optimal number of debate rounds varies with text length.

Limitations

Interaction Cost. D2D involves 5 debate stages and the coordination among 14 agents, resulting in considerable computational. To enable deployment in real-time settings, such as social media monitoring, future work may be expected to explore adaptive truncation strategies or lightweight models that maintain diversity without compromising quality.

Evidence Modality. Currently, D2D operates on textual input and does not incorporate external links, images, or videos, and thus lacks the capacity to detect multimodal misinformation such as deepfakes. Future work will focus on extending D2D's reasoning capabilities to encompass multimodal evidence, enabling more comprehensive misinformation detection.

Scalability and Real-time Adaptation. The performance of D2D is inherently tied to the capabilities of the underlying LLMs. Any deficiencies or biases in the LLM's pre-trained knowledge can propagate through the framework, affecting judgment reliability. This dependency introduces vulnerabilities, particularly when encountering domain-specific misinformation where LLM knowledge is outdated. Future work should consider integrating external knowledge bases, such as fact-checking repositories and domain-specific databases, to enhance real-time accuracy and reduce reliance on LLM-generated assumptions.

Ethics Statement

Our work aims to improve the interpretability and robustness of misinformation detection through the **Debate-to-Detect (D2D)** framework. We acknowledge the ethical considerations inherent in using LLMs for misinformation detection, particularly in environments such as political discourse and public health.

One major concern associated with the use of LLMs in misinformation detection is the potential for biased or erroneous inferences, which can arise from data imbalances or hallucinations. Furthermore, D2D's agent-driven debate, while designed to simulate human-like argumentation, may not fully replicate the nuance and contextual understanding required for real-world fact-checking, especially in culturally sensitive topics or politically charged narratives.

To address these challenges, our research implements several mitigation strategies. First, D2D applies domain-specific profiles to enhance the contextual relevance. Second, each agent is guided by structured prompts to maintain focus on evidence-based reasoning, minimizing the influence of speculative or unfounded claims. Finally, the judgment stage incorporates a multi-dimensional evaluation mechanism to assess the arguments, ensuring a balanced and transparent decision-making process.

Future work will also focus on enhancing bias detection and correction mechanisms, expanding domain coverage to include underrepresented topics, and incorporating external fact-checking databases to support more reliable judgment. Additionally, we plan to explore more robust fairness evaluations to improve the neutrality and reliability of D2D. We believe that our efforts to refine D2D's design and evaluation process are essential to mitigating the risks associated with LLM-driven misinformation detection, promoting fair and transparent deliberation across diverse information landscapes.

References

Alfonso Amayuelas, Xianjun Yang, Antonis Antoniades, Wenyue Hua, Liangming Pan, and William Yang Wang. 2024. MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6929–6948, Miami, Florida, USA. Association for Computational Linguistics.

Wang Angelina, Morgenstern Jamie, and Dickerson P. Dickerson. 2025. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. 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 (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why do multi-agent llm systems fail? *Preprint*, arXiv:2503.13657.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. KDD '24, page 6437–6447, New York, NY, USA. Association for Computing Machinery.

Rupak Kumar Das and Jonathan Dodge. 2025. Fake news detection after llm laundering: Measurement and explanation. *Preprint*, arXiv:2501.18649.

Zehang Deng, Yongjian Guo, Changzhou Han, Wanlun Ma, Junwu Xiong, Sheng Wen, and Yang Xiang. 2025. Ai agents under threat: A survey of key security challenges and future pathways. *ACM Comput. Surv.*, 57(7).

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. ICML'24. JMLR.org.

Aïmeur Esma, Amri Sabrine, and Brassard Gilles. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2698–2709, New York, NY, USA. Association for Computing Machinery.

Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9399–9416, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Huang, Jingyuan Yi, Peiyang Yu, and Xiaochuan Xu. 2025. Unmasking digital falsehoods: A comparative analysis of llm-based misinformation detection strategies. *Preprint*, arXiv:2503.00724.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7281–7294, Miami, Florida, USA. Association for Computational Linguistics.

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*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024a. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 3001–3004, New York, NY, USA. Association for Computing Machinery.

Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, and Diyi Yang. 2024b. Decoding susceptibility: Modeling misbelief to misinformation through a computational approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15178–15194, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *Preprint*, arXiv:2404.11584.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18798–18806.

M. F. Mridha, Ashfia Jannat Keya, Md. Abdul Hamid, Muhammad Mostafa Monowar, and Md. Saifur

Rahman. 2021. A comprehensive review on fake news detection with deep learning. *IEEE Access*, 9:156151–156170.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 3343–3347, New York, NY, USA. Association for Computing Machinery.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: Multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987, Miami, Florida, USA. Association for Computational Linguistics.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sougata Saha and Rohini Srihari. 2024. Integrating argumentation and hate-speech-based techniques for countering misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11109–11124, Miami, Florida, USA. Association for Computational Linguistics.

Upasna Sharma and Jaswinder Singh. 2024. A comprehensive overview of fake news detection on social networks. *Social Network Analysis and Mining*, 14(1):120.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Noam Slonim, Yonatan Bilu, Carlos Alzate, and 1 others. 2021. An autonomous debating system. *Nature*, 591:379–384.

Mubashir Sultan, Alan N. Tump, Nina Ehmann, Philipp Lorenz-Spreen, Ralph Hertwig, Anton Gollwitzer, and Ralf H. J. M. Kurvers. 2024. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proceedings of the National Academy of Sciences*, 121(47):e2409329121.

Spampatti Tobia, Hahnel Ulf J.J., Trutnevyte Evelina, and Tobias Brosch. 2024. Psychological inoculation strategies to fight climate disinformation across 12 countries. *Nature Human Behaviour*, 8:380–398.

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

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Shuai Xu, Jianqiu Xu, Shuo Yu, and Bohan Li. 2024. Identifying disinformation from online social media via dynamic modeling across propagation stages. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2712–2721, New York, NY, USA. Association for Computing Machinery.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *Preprint*, arXiv:2410.02736.

Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate. *Preprint*, arXiv:2408.04472.

Appendix

A Experiments on full datasets

In this appendix, we present the experimental results on the original datasets. Our main experiments are conducted on preprocessed versions of Weibo21 and FakeNewsDataset to ensure data quality and reduce noise. By evaluating the raw datasets, we illustrate how low-quality and ambiguous samples can degrade the model performance. The statistics of original datasets are presents in Table 6. The preprocessed datasets are also available at `4open.science/emnlp_d2d-36E2`.

A.1 Performance on Full Datasets

Table 7 presents the performance of the D2D framework on the original datasets. As observed, the results exhibit a decline compared to the preprocessed datasets, particularly in terms of Recall.

Dataset	Fake	Real	Average Words
Weibo21	2795	2956	92.08
FakeNewsDataset	490	490	276.12

Table 6: Statistics of two original datasets.

This disparity indicates that a significant portion of fake samples remains undetected by the model, resulting in a substantial number of false negatives.

A.2 Error Analysis

Upon analysis, a significant portion of the observed performance degradation can be attributed to low-quality samples, particularly in the Weibo21 dataset. These samples often exhibit poor structural coherence, substantial noise, or represent unverifiable claims that elude standard fact-checking procedures. We illustrate examples of these problematic samples in Figure 6. Consequently, the preprocessing is not only beneficial but necessary for enhancing model interpretability and performance consistency.

B Prompts Archive

Domain Inference:

Classify the domain of the following claim in one or two words (e.g., politics, finance, sports, technology, health). Claim:{input}

Profile Generation:

The domain is {domain}. Provide a brief professional profile (3-4 sentences) for a debater in {stage_name} stage role relevant to this domain.

Profile Example:

As a public health expert with over 15 years of experience, I have dedicated my career to ensuring the dissemination of accurate and evidence-based health information. I hold a Master's degree in Public Health from a leading university, supplemented by extensive research in health communication and media studies. My work has involved close collaboration with healthcare professionals, policy makers, and journalists to improve public understanding of health topics. This unique blend of expertise in both health and media allows me to effectively advocate for the crucial role factual news plays in shaping public health outcomes.

Shared Memory:

Given the following debate history: {debate_history}

Summarize the key points from both the Affirmative and Negative sides, ensuring the following aspects are preserved: 1. The main claim and its justification. 2. Key arguments and supporting evidence from both sides. 3. Notable rebuttals and counterarguments. 4. Any unresolved contradictions or logical conflicts.

Your summary should be concise yet comprehensive, allowing future agents to understand the debate's progression without losing important context. Aim to reduce redundancy while maintaining logical coherence.

Opening Statement:

{Profile}

The claim under discussion is: {input}. Your assigned stance is {fixed_stance}.

Based on your designated role and the available argument history, construct a well-structured opening statement that convincingly defends your stance. Make sure to employ logical reasoning, relevant evidence, and clear argumentation to support your position.

Rebuttal:

{Profile}

The claim under discussion is: {input}. Your assigned stance is {fixed_stance}. The previous argument presented was: {Shared_Memory}.

Identify the key weaknesses or logical inconsistencies in the opponent's argument and provide a well-structured rebuttal. Leverage relevant evidence and logical reasoning to effectively counter the claims made. Aim to challenge the validity of the argument while reinforcing your own position.

Free Debate:

{Profile}

The claim under discussion is: {input}. Your assigned stance is {fixed_stance}. The previous argument presented was: {Shared_Memory}.

Building on your previous arguments and responding to the latest claims, provide a well-structured continuation of the debate. Focus on addressing any unresolved contradictions, introducing new evidence if necessary, and strengthening your stance with logical reasoning.

Closing Statement:

{Profile}

The claim under discussion is: {input}. Your assigned stance is {fixed_stance}. The final evalu-

Method	Weibo21				FakeNewsDataset			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ZS	65.14	65.93	58.50	61.99	64.59	63.88	67.14	65.47
D2D	78.79	82.00	72.20	76.79	81.22	80.72	82.04	81.38

Table 7: Overall accuracy, precision, recall, and F1-score (%) on original datasets.

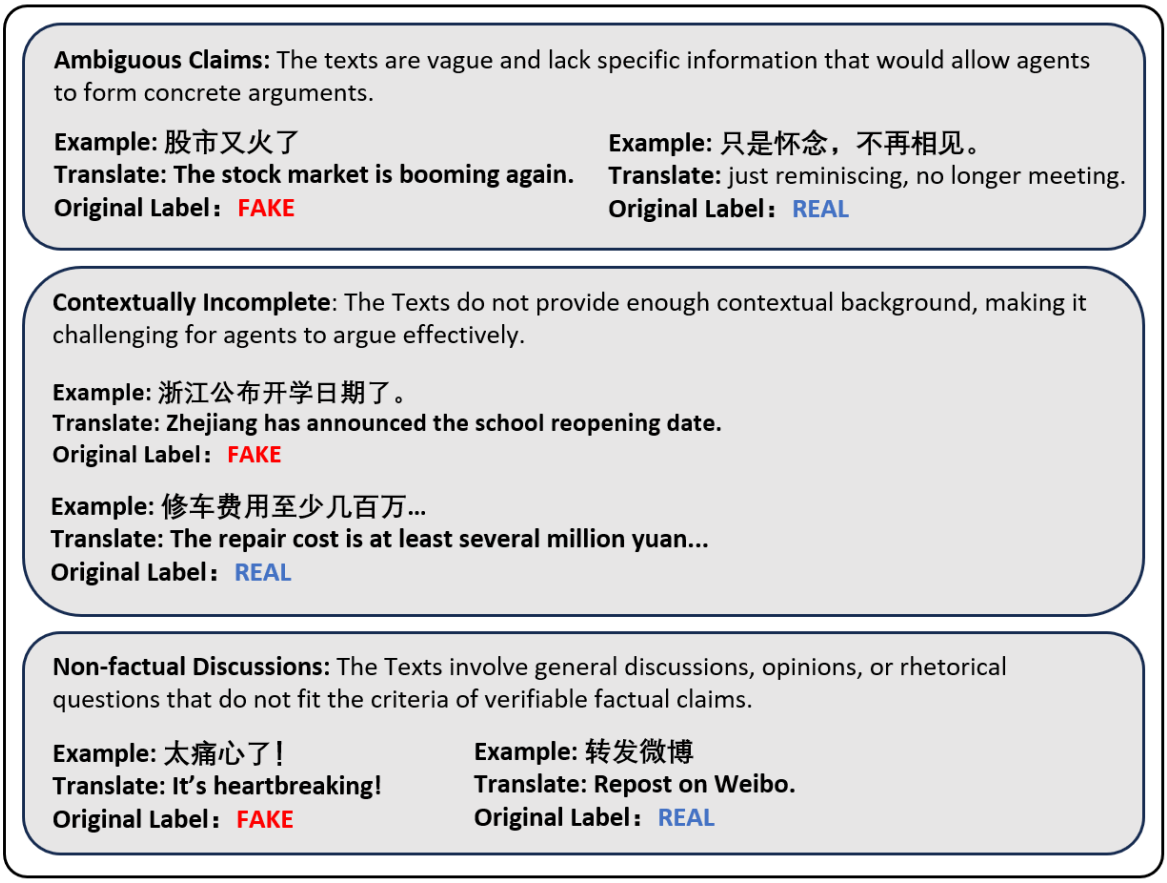

Figure 6: Examples of low-quality samples in Weibo21.

ation is approaching. The previous argument presented was: {Shared_Memory}.

Using this information, summarize your key arguments and highlight the most compelling evidence presented throughout the debate. Emphasize the logical coherence of your stance, address any lingering concerns or contradictions raised by the opposition, and consolidate your position. Conclude with a clear and decisive statement that reinforces your stance as the more rational and evidence-based perspective.

Judgement of Summary

{Profile}

You are assigned the role of a Judge responsible for summarizing the key points presented during the debate. Your task is to produce a concise and neutral summary that accurately reflects the main arguments from both the Affirmative and Negative sides.

The previous argument presented was: {Shared_Memory}.

Focus on the following aspects:

1. The main claim and its context.

2. Key supporting arguments presented by the Affirmative side.

3. Key counterarguments raised by the Negative side.

4. Notable rebuttals and their logical coherence.

5. Any unresolved contradictions or gaps in reasoning.

Judgement of Evaluation

{Profile}

You are assigned the role of a Judge, responsible for evaluating the quality and validity of the arguments presented during the debate. Affirmatives defend the claim as factual, and Negatives argue that the claim is misleading or fake.

The previous argument presented was: {Shared_Memory}.

Your task is to assess the arguments from both the Affirmative and Negative sides based on the {dimension_name} dimension.

For this dimension, assign an integer score to each side based on how convincingly they support their position relative to the truth. The two scores must add up to exactly 7.

Return the following JSON format:{Affirmative: X, Negative: Y}.