Deliberation-Based Multi-Agent Approach for Fallacy Detection

Anonymous ACL submission

Abstract

Detecting logical fallacies is essential for improving the quality of discourse and combating misinformation, yet current NLP methods are often hindered by the limitations of single-model analysis, which restricts reasoning diversity. Recent advancements in multi-agent systems using large language models (LLMs) have shown promise in tasks like fact-checking, but fallacy detection remains underexplored. In this paper, we introduce a new *deliberative multi-agent approach* for fallacy detection, where multiple LLM-based agents engage in structured deliberation to resolve their disagreements, mirroring ideal argumentation-theoretic practice. This approach improves reasoning rigor and allows for more subtle evaluations of fallacious arguments, outperforming single-LLM and standard ensemble methods, particularly in complex or ambiguous cases.

1 Introduction

The automated identification of logical fallacies has become an increasingly important task in natural language processing (NLP), due to its potential to mitigate misinformation and promote clearer, more rational discourse. Logical fallacies, arguments that appear persuasive despite containing underlying flaws, are pervasive in online discussions, political debates, and public communication (Jin et al., 2022). As such, developing robust methodologies for detecting fallacious reasoning is essential for advancing both critical thinking tools and trustworthy information systems.

Most existing approaches have focused on constructing benchmark datasets and developing classifiers, including large language models (LLMs), to identify fallacious arguments (Helwe et al., 2024; Jin et al., 2022). While these methods offer moderate success, they typically rely on a single model's judgment, limiting the diversity of reasoning and the potential for LLMs' interactive analysis.

Recent work has begun exploring multi-agent systems in which multiple LLMs interact, via debate or collaborative reasoning, to boost performance on tasks such as hallucination detection (Du et al., 2024), hate-speech detection (Park et al., 2024), and commonsense translation as well as arithmetic tasks (Liang et al., 2024). Yet, to our knowledge, no framework has been tailored specifically for fallacy detection. Moreover, most existing multi-agent studies emphasize debate-based interaction, leaving structured *deliberation* protocols largely unexplored, even though such protocols are crucial for rigorous and coherent evaluation. We present the first deliberative multi-agent approach explicitly designed for detecting logical fallacies.

Contributions. Our contributions are threefold: (1) we propose a deliberative multi-agent system for fallacy detection, in which multiple LLM-based agents engage in structured discussion to evaluate argument validity; (2) we compile the most comprehensive fallacy-detection benchmark to date and instantiate agents with a diverse suite of open-source LLMs (6B–70B parameters), enabling systematic analysis of model heterogeneity; and (3) we conduct extensive experiments against strong baselines to empirically demonstrate the effectiveness of our approach. All resources developed in this paper will be made publicly available.

2 Related Work

Fallacy Detection The automatic detection of logical fallacies in natural language has gained increasing attention in NLP.

Early work by Habernal et al. (2017) introduced ARGOTARIO, a game that both crowdsources and teaches fallacy recognition in order to gather annotated fallacious arguments at scale. This was followed by Habernal et al. (2018) where ad hominem argument types are mapped to their linguistic triggers. Goffredo et al. (2022) intro-

duces a 1,628-instance fallacy corpus covering six categories drawn from 31 U.S. presidential debates and shows that a transformer-based model leveraging argument components and relations surpasses previous baselines for fallacy classification. MAFALDA (Helwe et al., 2024) unifies prior fallacy datasets into a single benchmark with a harmonized taxonomy, supplies a manually annotated subset with explanations, introduces a subjectivity-aware annotation and evaluation framework, and reports zero-shot language-model and human performance on fallacy detection and classification. The LOGIC and LOGICCLIMATE datasets (Jin et al., 2022) frame logical-fallacy detection as a new reasoning benchmark, highlighting the task's value for both advancing NLP reasoning and curbing misinformation. Lei and Huang (2024) build an unsupervised logical-structure tree that captures connective-based reasoning and injects it into LLMs through hard (textual) and soft (embedding) prompts, improving precision and recall for both fallacy detection and classification. CoCoLoFa (Yeh et al., 2024) employ LLM-assisted crowdsourcing to create the largest fallacy dataset to date—7,706 annotated comments on 648 news articles—and shows that fine-tuning on this resource yields state-of-the-art detection and classification performance.

Multi-Agent Debate Systems Recent research shows that multi-agent debate systems enhance language model capabilities and improve reasoning and factuality by enabling viewpoints to challenge each other. Tree-of-Debate (Kargupta et al., 2025) transforms scientific papers into LLM personas that engage in a dynamically structured debate tree, yielding an expert-validated way to facilitate cross-domain literature review. Du et al. (2024) propose a "society-of-minds" framework in which multiple agents iteratively debate and converge on an answer, demonstrating a task-agnostic way to reduce hallucinations and enhance reasoning. PREDICT (Park et al., 2024) combines perspective-based reasoning and cross-stance debate to reconcile disparate labeling criteria and achieve state-of-the-art, cross-dataset hate-speech detection performance. Irving et al. (2018) propose a zero-sum self-play debate framework, in which two agents spar to reveal truthful information for a human judge, boosting accuracy scores and highlighting debate's promise and limitations for aligning AI with complex human goals. Liang et al. (2024) leverage judge-moderated debates to not only foster divergent reasoning but also mitigate the "Degeneration-of-Thought" in self-reflection. The research consistently demonstrates that structured disagreement between AI agents creates an emergent verification mechanism producing more reliable, factual, and creative outputs than traditional single-agent approaches.

3 Task and Data

Task Description We frame fallacy detection as a binary classification task, where each argument is labeled as either *fallacious* or *non-fallacious*. Each input instance consists of a complete argument, expressed at the sentence level or higher. The model's objective is to determine whether the reasoning presented in the argument contains a logical fallacy.

Data We create a unified *fallacy-detection* benchmark by merging the *test splits* of six publicly available datasets: MAFALDA (Helwe et al., 2024), CoCoLoFa (Yeh et al., 2024), LOGIC and LOGICCLIMATE (Jin et al., 2022), ElecDeb60To20 (Goffredo et al., 2023), and RuFal (Shultz, 2024). Because only MAFALDA and CoCoLoFa contain non-fallacious instances, we sample an additional 1,061 non-fallacious arguments from the CoCoLoFa *training* split to achieve class balance. The resulting benchmark comprises 2,882 arguments, exactly half fallacious and half non-fallacious, and spans multiple domains, including parliamentary debates, social-media threads, news comments, and climate-change articles. Table 1 shows the distribution of samples across sources. We partition the data into a 20 % validation set and an 80 % test set, using stratification on the class label to preserve the 50/50 balance in both subsets.

Datset	Count
CoCoLoFa	1859
LOGIC	300
MAFALDA	200
LOGICCLIMATE	219
ElecDeb60To20	199
RuFal	105

Table 1: Number of arguments per dataset.

4 Deliberation-Based Multi-Agent Framework

Unlike adversarial debate, which can reinforce polarization and binary framing, our approach em-

phasizes structured collaboration. We introduce a deliberative multi-agent framework in which multiple LLM-based agents engage in a cooperative reasoning process to assess whether an argument is fallacious. The goal is to foster nuanced judgment by encouraging agents to build on each other's reasoning rather than compete.

Deliberation proceeds in rounds. In each round, both agents provide an explanation of their reasoning and cast a vote: 0 for non-fallacious, 1 for fallacious. They must also output a confidence score in the range [0,1]. If the votes align, the classification is accepted and deliberation ends. If not, a new round begins in which the speaking order of the models is reversed. This ensures both models get a chance to respond to each other. This continues until consensus is reached or a maximum number of rounds is exceeded. If no agreement is reached, the vote with the higher confidence is selected; if confidence scores are equal, the label is chosen at random.

Agents follow a structured prompt emphasizing collaboration over confrontation. They can take four actions: *propose*, *argue*, *counter*, and *collaborate*. These encourage not only individual judgment but also engagement with the other agent's reasoning. The full deliberation prompt is included in Appendix B.

5 Experiments and Results

Baselines We evaluated six open-source LLMs from three language model families, each in two sizes: LLaMA 3.1 (8B Instruct and 70B Instruct), Mistral (7B Instruct and Small 24B 2501), and Command (Light 6B and 52B). Each model was prompted with three different techniques: zero-shot, few-shot, and Chain-of-Thought (CoT), yielding eighteen experimental conditions. Prompt details are provided in Appendix B. The highest accuracy scores on the validation set were achieved by LLaMA 70B in the zero-shot (70.2) and few-shot (70.0) settings. The test results are reported in Table 2 whereas the validation results can be found in Table 4 in Appendix A.

Agreement-based Ensemble Strategy We integrate the two top-performing configurations by merging their predictions. When both models agree on a classification, their shared prediction is taken as the output. In cases of disagreement, the models engage in deliberation within our *framework* to reach a consensus. This layered strategy ensures that additional computation and token usage are reserved for difficult cases, while straightforward instances are resolved efficiently. For comparison, we include an alternative resolution method that randomly selects one of the two conflicting labels.

	LLaMA		**Mistral**		**Command**	
	8B	70B	7B	24B	6B	52B
Zero	65.0	**69.9**	58.0	56.4	46.2	62.2
Few	67.0	**71.9**	65.2	66.6	48.9	58.8
CoT	60.5	**68.9**	64.9	67.3	55.4	53.6

Table 2: Accuracy scores on the test set (n=2305). Zero: zero-shot; Few: few-shot; CoT: Chain-of-Thought.

Evaluation The original dataset labels are mapped to binary classes: *fallacious* (1) or *non-fallacious* (0). We report accuracy on the balanced validation and test sets. For the subset of instances where the top-performing models disagree (i.e., those requiring deliberation) we report macro F1-score, as the class distribution in this subset is imbalanced. During inference, models are instructed to output either `0` or `1`. In the deliberation setting, models must include a `<vote>` tag in their response to indicate the final classification. Any output that cannot be mapped to one of the binary labels is treated as invalid and counted as incorrect. In the ensemble strategy using the two best-performing models, a prediction is marked incorrect only if *both* outputs are invalid. If only one model produces an invalid output, the valid prediction from the other model is accepted as the final label.

Results Table 3 presents the test set performance for the ensemble methods alongside the two strongest baselines. The best individual result was achieved by the LLaMA 70B model with few-shot prompting, reaching an accuracy of 71.9%. The ensemble of the top two models with random tie-breaking slightly underperformed, yielding 70.8% accuracy. The highest accuracy, 72.4%, was obtained using our deliberative resolution strategy.

Of the 2,305 test instances, the two strongest models agreed on 1,875 cases, achieving an accuracy of 75.8%. Disagreements occurred on 416 instances, with an additional 14 cases excluded due to unparsable outputs. Following one round of deliberation, 408 disagreements were resolved through consensus, 3 were decided based on a single valid output, and the remaining 5 were resolved via random selection. On the 416 disputed instances, the

	Accuracy
Zero-shot	69.9
Few-shot	71.9
Ensemble + random	70.8
Ensemble + deliberation	**72.4**

Table 3: Test set accuracy scores for the two strongest baseline methods and the ensemble method using two different resolution strategies: random selection and our deliberation framework.

deliberation achieved a macro F1-score of 57.2.

Discussion The strongest baseline is LLaMA 70B, which is also the largest model evaluated. Within each model family, larger variants consistently outperform their smaller counterparts, indicating a clear model size effect. However, this trend does not generalize across model families: LLaMA 8B outperforms both Cohere models in all configurations and surpasses Mistral in all but the CoT prompting setting. These results suggest that model architecture and training data (i.e., the model family) have a greater impact on performance than size alone. This finding is encouraging for efforts to reduce the computational demands of LLMs without sacrificing accuracy.

Zero-shot performance for LLaMA is already strong, especially considering the absence of task-specific tuning. Few-shot prompting improves performance for Mistral, but has limited or no effect on Cohere models. CoT prompting produces mixed results: it enhances performance for Mistral 7B and Cohere 6B, but leads to a decline for LLaMA 8B and Cohere 52B. Also, CoT prompts introduced significant formatting inconsistencies for Command Light, requiring an adjusted prompt with stricter output guidelines.

Our ensemble strategy combines the two best-performing baselines to use their complementary strengths while minimizing unnecessary computation on easy cases. However, applying simple random tie-breaking slightly degrades performance, likely due to the noise introduced by arbitrary decisions on difficult instances. The best overall performance is achieved through structured deliberation, yielding a modest improvement over the strongest individual model. This demonstrates the effectiveness of targeted resolution strategies for ambiguous or contentious cases. Notably, the deliberation accuracy on the 416 disagreement instances is 57.2%, highlighting both the intrinsic difficulty of these examples and the potential for further gains through more advanced reasoning mechanisms.

The deliberations are often of high quality, with agents engaging constructively, building on one another's points, and demonstrating an awareness of ambiguity (e.g., "I acknowledge that others may interpret the argument differently"). However, most disagreements were resolved within a single round, suggesting an over-eagerness to reach consensus, potentially at the expense of deeper critical engagement. Manual analysis of a small sample revealed that agents frequently layered points, such as agreeing and then adding an additional consideration, which may inhibit direct rebuttal and limit adversarial challenge. While deliberation promotes collaborative reasoning, it may lack the adversarial dynamics necessary for sharper discrimination. A hybrid strategy that uses deliberation for exploration and structured debate for resolution could offer a more effective approach, particularly in borderline or ambiguous cases. Formatting errors occurred in fewer than 1% of cases and could be mitigated by constraining outputs to structured formats. However, prior work suggests that such constraints may lead to performance degradation by limiting model expressiveness. Finally, the confidence scores, intended as a fallback mechanism to resolve deadlocks when the maximum number of deliberation rounds is exceeded, proved largely uninformative. In over 90% of cases, models output values of 0.8 or 0.9, reflecting an overconfident bias and limiting their usefulness for reliable conflict resolution.

6 Conclusion

We propose a deliberative multi-agent approach for logical fallacy detection that enables structured collaboration between language models. Our approach outperforms both single-agent baselines and an ensemble strategy. For evaluation, we introduce a unified benchmark by aggregating six existing fallacy detection datasets and conduct extensive experiments across eighteen prompting configurations, using diverse open-source LLMs. While larger models within the same architecture tend to perform better, differences in model family have a greater impact than size alone, as shown by LLaMA 8B outperforming much larger models. Our two-step strategy is compute-efficient: agreement is reached on over 80% of instances, reserving deliberation for harder cases. This balances performance with reduced computational cost.

Limitations

Most disagreements were resolved in a single round, indicating a tendency for quick agreement that may hinder a more rigorous analysis. While the deliberative setup improved collaborative reasoning, it may benefit from a more adversarial tone to encourage a deeper critical engagement with the arguments.

Ethical Statement

This paper addresses the task of detecting logical fallacies, flaws or errors in reasoning that can appear across various forms of human communication. Such fallacies may contribute to the spread of misinformation or reinforce societal biases, leading to harmful consequences. The primary goal of this research is to better understand and identify logical fallacies in natural language.

References

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. Argument-based detection and classification of fallacies in political debates. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. MAFALDA: A benchmark and comprehensive study of fallacy detection and classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *Preprint*, arXiv:1805.00899.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. 2025. Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis. *Preprint*, arXiv:2502.14767.

Yuanyuan Lei and Ruihong Huang. 2024. Boosting logical fallacy reasoning in LLMs via logical structure tree. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13157–13173, 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.

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.

Benjamin Shultz. 2024. An entity-aware approach to logical fallacy detection in kremlin social media content. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '23, page 780–783, New York, NY, USA. Association for Computing Machinery.

Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-

assisted crowds. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.

A Validation Results

	LLaMA		Mistral		Command	
	8B	70B	7B	24B	6B	52B
Zero	65.7	**70.2**	57.5	56.2	48.9	60.7
Few	67.1	**70.0**	64.5	66.7	51.8	58.9
CoT	62.0	**69.0**	65.9	67.2	51.1	57.5

Table 4: Accuracy scores on the validation set (n = 577). Zero: zero-shot; Few: few-shot; CoT: Chain-of-Thought.

B Prompts

For models that do not support system prompts, the system instructions were prepended to the first user message.

B.1 Zero-shot

<system>Always output a 0 or 1. Do not say anything else.</system>

<user>Instructions: Output a "1" if the argument is fallacious, otherwise "0". Argument: [argument]</user>

B.2 Few-shot

<system>Always output a 0 or 1. Do not say anything else.</system>

<user>Instructions: Output a "1" if the argument is fallacious, otherwise "0". Argument: Obviously, we elect evil and allow Fabricio to eliminate the rights of those he does not agree with or we vote for Carlos and guarantee equal rights for all. Or are there more options for election? Is there anyone who is more moderate on all issues?</user>

<assistant>1</assistant>

<user>Instructions: Output a "1" if the argument is fallacious, otherwise "0". Argument: So no matter what they think, so long as most people believe in something, it's okay? Most people thought slavery was okay at some point too. Not everything can be left up to public opinion of the time. The protection of these gullies is extremely important to future generations and more should be done to protect them, even if most people are maybe ignorant or short-sighted about the harm they're really causing.</user>

<assistant>0</assistant>

<user>Instructions: Output a "1" if the argument is fallacious, otherwise "0". Argument: [argument]</user>

B.3 Chain-Of-Thought

B.3.1 General

<system>You are a reasoning assistant trained to detect logical fallacies. Follow this reasoning chain before answering:

1. Restate the argument in your own words.
2. Identify the conclusion and the supporting premises.
3. Examine whether the conclusion logically follows from the premises.
4. Check if the argument fits any known fallacy types (e.g., ad hominem, straw man, slippery slope, false dilemma, circular reasoning, appeal to authority/popularity, etc.).
5. If a fallacy is present, output 1.
6. If the reasoning is sound and no fallacy is detected, output 0.

Always output a 0 or 1. Do not say anything else.</system>

<user>Instructions: Output a "1" if the argument is fallacious, otherwise "0". Argument: [argument]</user>

B.3.2 Command Light

<user>You are a reasoning assistant trained to detect logical fallacies. Follow this step-by-step reasoning process internally before producing your final output:

- Restate the argument in your own words.
- Identify the conclusion and the supporting premises.
- Examine whether the conclusion logically follows from the premises.
- Check if the argument fits any known fallacy types (e.g., ad hominem, straw man, slippery slope, false dilemma, circular reasoning, appeal to authority/popularity, etc.).

Decision Rule: If a fallacy is present, output "f". If no fallacy is present and the reasoning is sound, output "v".

Output Rules
Your final response must be ONLY a single character: v or f. Do not write any words, explanations, or punctuation. Just output v or f.

Argument: [argument]</user>

B.4 Deliberation

<system>## Instructions
You are a Deliberator Agent in a structured, high-quality reasoning process. Your task is to evaluate whether a given argument contains a fallacy or not. The goal is to make careful, well-supported, and collaborative judgments about the soundness of arguments.

Actions

- Propose — Clearly state your initial judgment: is the argument fallacious or not?
- Argue — Justify your assessment using critical thinking. Analyze the argument's structure, relevance, and validity. Engage critically with the arguments to go beyond surface-level relevance.
- Counter — Respond to others' assessments. Consider ambiguity, provide clarifications, or refine your own claim.
- Collaborate — Work toward consensus. Compare interpretations and weigh competing perspectives constructively.

Guidelines
Be charitable: Most arguments are imperfect. Do not label an argument fallacious unless there is a clear breakdown in reasoning.

Fallacies must affect validity: Only identify a fallacy if it undermines the argument's logical structure or reasoning.

Favor precision over pedantry: Focus on argument quality, not on language or phrasing quirks unless they cause confusion.

Respect human annotation standards: Human interpretation is the gold standard. Keep judgments reasonable and aligned with real-world argumentation.

Output format
First output your contribution to the deliberation, do not use the <deliberation> tags, then:
if you think the argument is fallacious, output: <vote>1</vote>
If you think the argument is valid, output: <vote>0</vote>
Additionally, output your confidence score (0-1) which expresses how certain you are in your vote: <confidence>[0-1]</confidence><system> <user>Here is the current state of the deliberation: <deliberation>[context]</deliberation>
Based on the discussion above, continue the deliberation and cast your vote.</user>