Multi-Agent LLM Debate Unveils the Premise Left Unsaid

Harvey Bonmu Ku[1,2] **Jeongyeol Shin**[2] **Hyoun Jun Lee**[2] **Seonok Na**[2] **Insu Jeon**[2]
[1]Ministry of National Defense, Republic of Korea
[2]Qraft Technologies
language@langua.ge
{jeongyeol.shin, hyounjun.lee, seonok.na}@qraftec.com
insuj3on@gmail.com

Abstract

Implicit premise is central to argumentative coherence and faithfulness, yet remain elusive in traditional single-pass computational models. We introduce a multi-agent framework that casts implicit premise recovery as a dialogic reasoning task between two LLM agents. Through structured rounds of debate, agents critically evaluate competing premises and converge on the most contextually appropriate interpretation. Evaluated on a controlled binary classification benchmark for premise selection, our approach achieves state-of-the-art accuracy, outperforming both neural baselines and single-agent LLMs. We find that accuracy gains stem not from repeated generation, but from agents refining their predictions in response to opposing views. Moreover, we show that forcing models to defend assigned stances degrades performance—engendering rhetorical rigidity to flawed reasoning. These results underscore the value of interactive debate in revealing pragmatic components of argument structure.

1 Introduction

Arguments do not fail at the surface; they often fail in what they assume. What makes an argument persuasive is not always what is stated, but what is left unsaid. Implicit premises—unstated assumptions that connect reasons to claims—are often the true engines of argumentation (Hitchcock, 1985; Toulmin, 1958; Walton and Reed, 2005).

Recovering implicit premises thus represents a foundational, yet underexplored, challenge in computational argument analysis. Existing systems perform well at identifying explicit argumentative components such as claims and reasons, but they often fall short in capturing what is pragmatically presupposed (Feng and Hirst, 2011; Walton and Reed, 2005; Habernal et al., 2018a).

This limitation becomes particularly consequential in high-stakes domains such as law, finance, and politics, where arguments frequently hinge on assumptions that are unstated, ambiguous, or implied (Chakrabarty et al., 2021). In such contexts, argument mining must move beyond surface-level interpretation to reconstruct the hidden connective tissue that underpin argumentative coherence (Hitchcock, 1985; Razuvayevskaya and Teufel, 2017; Katz et al., 2022).

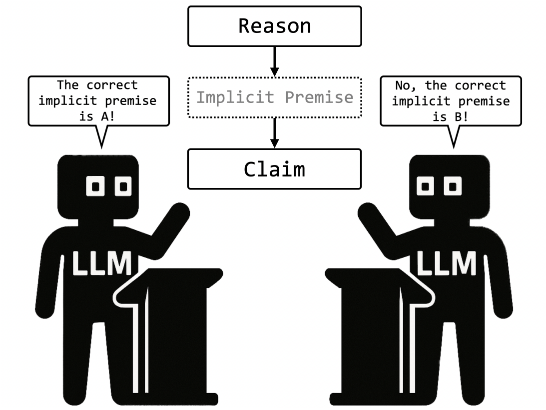

Figure 1: Illustration of the two LLM agents debating which is the correct implicit premise.

The advent of large language models (LLMs) has opened new possibilities for modeling contextual reasoning at scale. Yet when applied to tasks demanding pragmatic inference, LLMs operating in isolation often fall short (Katz et al., 2022; Chakrabarty et al., 2021). A key limitation is their inability to interrogate their own outputs; reflective techniques such as self-reflection (Shinn et al., 2023) are often unsuitable for capturing the nuanced reasoning required in argument mining. In natural discourse, implicit premises are rarely surfaced in isolation—they are negotiated through interaction, clarification, and iterative exchange (Inoue et al., 2020; Stede et al., 2019).

Motivated by this observation, we propose a multi-agent framework that models premise recovery as a dialogic reasoning process between two LLM agents. This approach draws on recent findings that language models demonstrate more co-

Proceedings of the 12th Argument Mining Workshop, pages 58–73
July 31, 2025 ©2025 Association for Computational Linguistics

herent reasoning in interactive settings (Du et al., 2024), and show enhanced pragmatic sensitivity when engaged in debate (Ku, 2025). In our setup, agents are either assigned or select a candidate premise and interact either sequentially or simultaneously through structured rounds of deliberation.

We evaluate this method on the SemEval 2018 Task 12 dataset (Habernal et al., 2018b), which casts implicit premise recovery as a binary classification task. While prior models—including LSTM and BERT-based classifiers—showed moderate success, our multi-agent approach achieves the highest accuracy to date, outperforming both traditional baselines and single-agent LLMs. These results underscore the potential of agentic reasoning as a more effective paradigm for capturing the pragmatic inference required in implicit argument understanding. The primary contributions of this work are as follows:

- We position implicit premise recovery as a central task in argument mining, moving beyond surface-level extraction toward modeling the pragmatic reasoning that underlies argumentative coherence.

- We propose a multi-agent LLM framework that addresses premise selection as a dialogic process, yielding state-of-the-art performance on a benchmark dataset.

2 Related Work

2.1 Implicit Premises and Deeper Argument Understanding

The task of recovering implicit premises—unstated assumptions that bridge claims and reasons—is closely related to enthymeme reconstruction in classical argumentation theory. Enthymemes omit one or more components of an argument, typically leaving the audience to infer missing premises. Recovering these implicit links is crucial for argument mining, as they often carry the inferential burden behind persuasive discourse. Early work highlighted the logical challenges of modeling enthymemes (Hitchcock, 1985), while more recent studies have focused on detecting, classifying, or generating missing premises (Boltužić and Šnajder, 2016; Rajendran et al., 2016; Chakrabarty et al., 2021; Hunter, 2022; Stahl et al., 2023).

Building on this line of inquiry, researchers have investigated a range of tasks that involve implicit inference, including the recovery of unstated reasoning chains in question answering (Katz et al., 2022), the identification of event arguments with long-range dependencies (Lin et al., 2022a), and the discovery of relational links between argumentative units via implicit inferences (Saadat-Yazdi et al., 2023). These studies show that even state-of-the-art systems often struggle to model the background knowledge and pragmatic logic required to make sense of incomplete arguments.

Beyond model development, recent efforts have sought to improve the quality of annotated data for implicit reasoning. Singh et al. (2021) proposed a semi-structured annotation methodology for collecting implicit warrants, demonstrating that abstract assumptions can be reliably captured via guided crowdsourcing.

While these advances have expanded our understanding of hidden argumentative structure, implicitness is still often treated as a supporting concern rather than a central modeling objective. In contrast, our work foregrounds implicit premise recovery as the primary task and frames the process as one of pragmatic, dialogic reasoning between agents.

2.2 Multi-Agent LLM Debate

Multi-agent debate has emerged as a promising method for enhancing reasoning in large language models by transforming inference from a solitary act into an interactive process. Instead of relying on a single model's output, multiple agents engage in dialogue—critiquing, revising, and refining their interpretations—mirroring the deliberative nature of human reasoning (Irving et al., 2018; Du et al., 2024). Such interactions improve factual accuracy, consistency, and interpretability across domains. Chan et al. (2024) and Liang et al. (2024) report that multi-agent discussions help overcome individual model biases, with Liang et al. (2024) describing this as a remedy for the "degeneration-of-thought" effect—where flawed lines of reasoning persist without external correction. These insights echo Minsky (1988)'s notion of a "society of minds," in which intelligence arises from the interplay of multiple specialized reasoning units.

We extend this paradigm to the domain of argument mining, where implicit premise recovery requires more than the injection of external knowledge—it demands interpretive contrast. To our knowledge, this is the first study to apply multi-agent LLM debate to an argument mining task.

3 Methodology

3.1 Task Definition

We define the task of *implicit premise recovery* as selecting the correct implicit premise $P^* \in \{\text{Premise A}, \text{Premise B}\}$ that logically and pragmatically bridges a reason R and a claim C in a given argument tuple $x = (C, R, \text{Premise A}, \text{Premise B})$.

Claim:	Young people's votes matter.
Reason:	All votes matter.
Premise A:	Many young people vote.
Premise B:	Many young people don't vote.

Table 1: Example of an implicit premise recovery instance.

This example highlights the subtlety of the task: both candidate premises appear logically plausible yet imply distinct pragmatic interpretations. Premise A implies descriptive inclusion—that young people are already voters whose contributions merit recognition—while Premise B suggests normative urgency, highlighting that their underrepresentation makes their votes especially valuable. Disambiguating between such readings requires sensitivity to context and intent, rather than reliance on lexical overlap or surface logic.

We approach this task as a deliberative process between two large language model agents, each initialized with a different candidate premise. Through structured multi-round dialogue, the agents attempt to resolve their disagreement and identify the premise P^* that most plausibly completes the argument.

Formally, a debate instance D consists of a sequence of rounds $D = \{R_1, R_2, \ldots, R_n\}$, where each round R_i contains contributions $(a_i^{(A)}, a_i^{(B)})$ from agents A and B, respectively. The task is evaluated as a binary classification problem: each instance is marked correct if the final agreed-upon premise matches the gold-standard label, or incorrect if the debate either results in the wrong selection or terminates without consensus after n rounds.

3.2 Design of the LLM Debate

To systematically evaluate how LLMs reason over competing premises, we design a debate framework that manipulates two key structural conditions: stance assignment and interaction order. These conditions allow us to test how different configurations affect argumentative convergence and overall performance.

Condition 1: Given vs. Chosen Stance In the *Given* stance condition, each agent is explicitly assigned a candidate premise to defend—either Premise A or Premise B. During preliminary testing, we observe that agents often rigidly maintain their initial stance, even when logically weaker (see Appendix Figure 8). To address this, we introduce staged prompting: early rounds emphasize advocacy, while later rounds prompt agents to neutrally evaluate both premises and converge on the more plausible one (Appendix Listing 2).

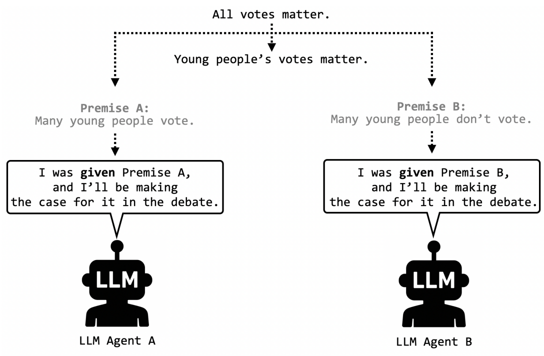

Figure 2: Illustration of the Given conditon.

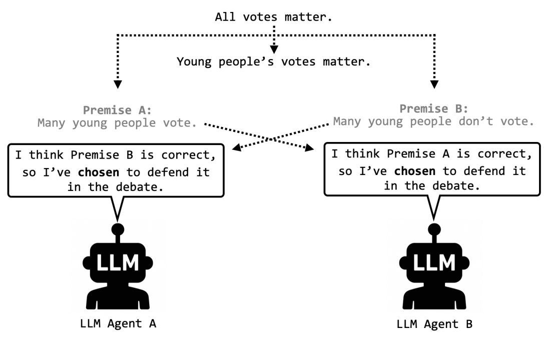

Figure 3: Illustration of the Chosen conditon.

In the *Chosen* stance condition, each agent independently selects the premise it finds more convincing and is instructed to defend that choice (Appendix Listing 3). If the agents agree on a premise early in the debate, the session is immediately terminated and the shared answer is evaluated against the gold label.

Condition 2: Sequential vs. Simultaneous Round While the first round in both configurations functions as an opening statement—analogous to initial remarks in formal debate—the two conditions diverge in how subsequent rounds are structured and processed.

In the *sequential* setup, agents engage in alternating turns; Agent A begins by defending one candidate premise, and Agent B responds after reviewing A's output. This allows each agent to build on or challenge the preceding argument.

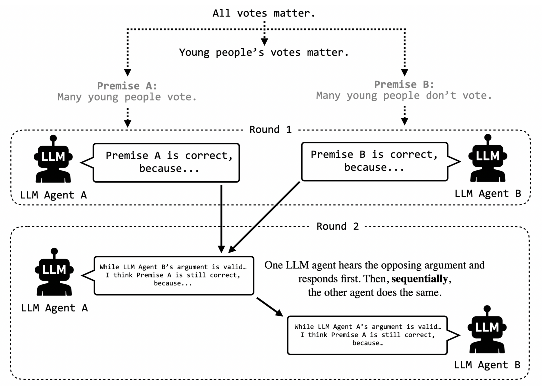

Figure 4: Illustration of the Sequential conditon.

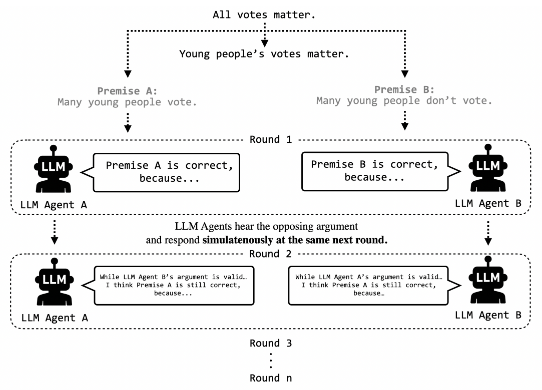

Figure 5: Illustration of the Simultaneous conditon.

In the *simultaneous* setup, both agents produce their arguments independently and then respond to each other's initial outputs in the following round. This structure enables a more parallel and symmetrical form of interaction.

This design allows us to evaluate whether free premise selection improves convergence and whether agents benefit from observing each other's arguments across rounds. The full implementation details, including model selection, decoding parameters, and logging tools, are provided in the subsequent section.

4 Experimental Setup

4.1 Dataset

We evaluate our approach using the Argument Reasoning Comprehension Task dataset from SemEval-2018 Task 12 (Habernal et al., 2018b), a benchmark explicitly designed to test implicit reasoning in natural language arguments. Each instance consists of a claim, a reason, and two candidate warrants[1]—only one of which correctly links the reason to the claim.

The incorrect premises are crafted to be topically and lexically plausible, yet logically incompatible with the argument, thereby requiring models to engage in pragmatic inference rather than rely on shallow surface cues. The dataset contains 1,970 instances drawn from online debates, partitioned into training (1,210), development (316), and test (444) sets. For our evaluation, we focus on the held-out test set to enable direct comparison with previously reported results from baseline models.

This dataset is particularly well-suited for our purposes because of (1) its topical diversity—including politics, ethics, economics, and social policy—which mirrors real-world argumentative variety, and (2) its construction via a rigorous eight-step crowdsourcing pipeline with multiple validation rounds, ensuring that examples are high-quality and pragmatically meaningful.

4.2 Model Configuration

We implement all experiments using OpenAI's GPT-4o-mini, the most cost-effective and fastest available LLM at the time of writing. Given the latency introduced by multi-turn agent interaction, GPT-4o-mini offers the best balance between computational efficiency and linguistic performance. All LLM experiments—including the single-agent baseline—use identical model settings to ensure comparability. Multi-agent interactions are managed via the LangGraph framework, which facilitates node-based orchestration and message passing. Logging and analysis of outputs are performed using LangSmith.

4.3 Parameters

To determine appropriate parameters, we conducted preliminary experiments using the single-agent LLM. We tested temperature values of 0.1, 0.3, 0.5, 0.7, and 0.9, along with max round settings of 5, 10, 15, and 20. Neither parameter showed statistically significant impact on performance. We therefore adopted the median configuration: tem-

[1] We treat "warrant" and "implicit premise" interchangeably throughout this paper, following Toulmin's framework (Toulmin, 1958) in which a warrant serves as the unstated bridge in an argument.

perature was fixed at 0.5, and all debates were capped at 10 rounds.

No few-shot examples or chain-of-thought prompting were used. Given that implicit premise recovery is a pragmatic reasoning task with no canonical steps, such scaffolding was treated as a potential confound. If no agreement was reached within 10 rounds, the debate was marked incorrect.

4.4 Previous Models

To establish strong baselines for comparison, we replicated two representative models for implicit premise recovery. Rather than relying solely on reported metrics, we reproduced both models using their publicly available code and the original test dataset.

LSTM This model designed by Choi et al., 2018 implements a hybrid architecture combining a pre-trained Enhanced Sequential Inference Model (ESIM; Chen et al., 2017) with a bidirectional LSTM. The ESIM component, trained on SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), captures entailment knowledge and passes frozen sentence pair representations to a task-specific BiLSTM. The model processes all relevant pairings—claim–premise, premise–reason, and premise–premise—and feeds their concatenated outputs into a fully connected network to determine the correct implicit premise. This approach ranked first in the 2018 shared task and outperformed all other submissions by a margin of over 10 percentage points (Habernal et al., 2018b).

BERT We fine-tuned RoBERTa (Liu et al., 2019), an optimized variant of BERT that omits the Next Sentence Prediction objective and is trained on longer sequences and larger corpora. Inputs were formatted as concatenated sequences of the claim, reason, and candidate implicit premise. Compared to sequential models like LSTM, RoBERTa uses self-attention to capture contextual dependencies across the entire input simultaneously. The model was trained for 10 epochs with a learning rate of $1e^{-5}$, weight decay of 0.01, and a batch size of 16. The maximum sequence length was set to 512 tokens, and all experiments were run on 8 A100 GPUs.

5 Results

5.1 Main Results

Table 2 presents a comparison of model performance across prior baselines and the five LLM configurations tested in this study. The single-agent LLM baseline achieved an accuracy of 0.7928, outperforming previous neural models—including the top-performing LSTM (0.7050) and a fine-tuned RoBERTa model (0.7564). This result confirms that a single-pass LLM does exhibit strong capabilities for implicit premise recovery under zero-shot conditions.

Our multi-agent framework, however, produced further improvements under specific configurations. In *Chosen* stance setups—, where agents independently selected and defended their preferred premise—, both interaction orders led to substantial gains. The *Simultaneous* condition achieved 0.8446 in accuracy, and the *Sequential* condition yielded the highest overall performance at 0.8694. These results indicate that dialogic reasoning is most effective when agents are free to align on a shared interpretation, rather than being constrained by initial position assignments.

A Cochran's Q test confirmed a statistically significant difference in performance across the five LLM configurations ($Q = 101.03$, $df = 4$, $p < 0.0001$), prompting further pairwise analysis. Post-hoc McNemar tests revealed that nearly all model pairs differed significantly, with two key exceptions. First, the two highest-performing conditions—*Chosen & Sequential* and *Chosen & Simultaneous*—did not differ significantly ($p > 0.05$), despite a nominal accuracy gap of 2.5 percentage points. Second, the two lowest-performing configurations—*Given & Sequential* and *Given & Simultaneous*—also showed no significant difference ($p > 0.05$), suggesting that interaction order exerted limited influence in the presence of fixed stance assignments.

Direct comparisons with the single-agent baseline further clarify this pattern. The single-agent LLM statistically outperformed both *Given* stance conditions: for *Given & Sequential*, the McNemar test yielded $p < 0.0001$ (contingency: 275 both correct, 77 single only, 35 *Given* only, 57 both wrong); for *Given & Simultaneous*, $p < 0.01$ (contingency: 288, 64, 32, 60). These results indicate that rigid stance assignment may suppress performance even relative to non-interactive inference.

Conversely, both *Chosen* stance configurations

Model	Citation	Accuracy	Precision	Recall	F1
		Previous Studies			
Baseline	Habernal et al. (2018b)	0.5000	-	-	-
LSTM	Choi et al. (2018)	0.7050	0.7281	0.6870	0.7069
BERT	Liu et al. (2019)	0.7564	0.7568	0.7568	0.7568
		LLM-based experiments			
Single-agent LLM	This study	0.7928	0.7941	0.7928	0.7928
MultiAgent Debate *(Given & Sequential)*	This study	0.6982	0.6986	0.6982	0.6973
MultiAgent Debate *(Given & Simultaneous)*	This study	0.7207	0.7207	0.7207	0.7207
MultiAgent Debate *(Chosen & Sequential)*	This study	**0.8694**	**0.8768**	**0.8694**	**0.8691**
MultiAgent Debate *(Chosen & Simultaneous)*	This study	0.8446	0.8553	0.8446	0.8440

Table 2: Comparison of performance on implicit premise recovery across prior models and configurations tested in this study. The best scores are in bold.

significantly outperformed the single-agent model. Against *Chosen & Sequential*, the McNemar test yielded $p < 0.0001$ (contingency: 337, 49, 15, 43); against *Chosen & Simultaneous*, $p < 0.01$ (contingency: 334, 41, 18, 51). These findings confirm that when agents are permitted to self-select and defend their preferred stance, multi-agent interaction leads to robust improvements over single-pass prompting.

Taken together, these results indicate that stance assignment—not interaction order—is the primary determinant of performance differences in multi-agent LLM debate. While alternating turns may allow for richer back-and-forth refinement, its impact is modest compared to the benefits of allowing agents to converge on shared, self-selected premises.

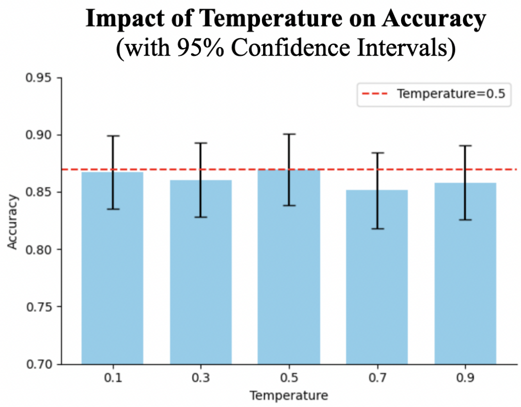

Figure 6: Impact of decoding temperature on implicit premise recovery accuracy under the *Chosen & Sequential* setting. Error bars represent 95% confidence intervals.

5.2 Effect of Temperature and Max Rounds

To test whether decoding parameters affect performance, we conducted an additional set of experiments using the best-performing configuration—*Chosen & Sequential*—as a base. While this setting yielded the highest overall accuracy (0.8694), it was not statistically distinguishable from the *Chosen & Simultaneous* condition ($p > 0.05$), indicating that both settings perform comparably under the chosen evaluation metric.

We varied the temperature parameter across five values (0.1, 0.3, 0.5, 0.7, and 0.9), holding all other factors constant. Temperature 0.5 was used throughout our main experiments, including both single-agent and multi-agent runs. A Cochran's Q test revealed a highly significant difference across the five temperature conditions ($Q = 150.18$, $df = 4$, $p < 0.0001$), suggesting that temperature meaningfully impacts model behavior at the instance level—even when overall accuracy remains comparable (ranging from 0.8514 to 0.8694). Post hoc McNemar tests confirmed that temperature 0.5 differs significantly from all other settings: 0.1 ($p < 0.0001$), 0.3 ($p < 0.0001$), 0.7 ($p < 0.0001$), and 0.9 ($p < 0.0001$). In contrast, no significant differences were observed between any of the non-0.5 pairs. These findings indicate that temperature 0.5 produces a statistically distinct profile of correct predictions while yielding the highest accuracy among tested settings.

To examine whether the number of debate rounds influences performance, we conducted a similar test across four configurations (N = 5, 10, 15, 20). A Cochran's Q test yielded no significant difference across these settings ($Q = 0.063$, $df = 3$, $p > 0.05$), suggesting that extending or shortening the debate window has minimal effect on instance-level behavior. Accordingly, we re-

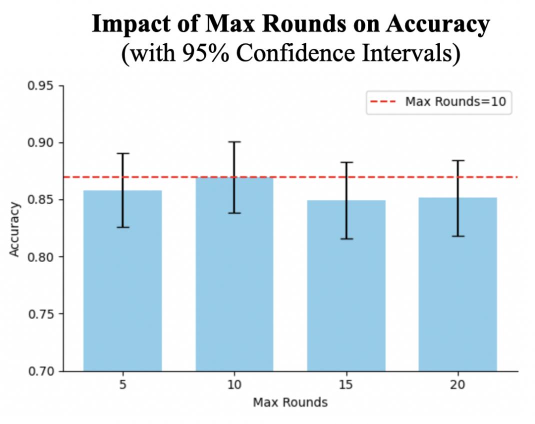

Figure 7: Impact of maximum number of debate rounds on accuracy under the *Chosen & Sequential* setting. Error bars represent 95% confidence intervals.

tain $N = 10$ as a reasonable and computationally efficient default for all primary experiments.

6 Discussion

6.1 Effectiveness of Multi-Agent Debate

Our multi-agent debate framework outperforms all previous models—surpassing LSTM-based systems, fine-tuned BERT classifiers, and single-agent LLMs—on the task of implicit premise recovery. Crucially, this improvement is not merely an artifact of increased generation length or system complexity. Rather, we argue that performance gains arise because agents iteratively refine their beliefs in response to alternative perspectives, producing more robust and context-sensitive inferences (as evidenced in Appendix Figure 12).

One may reasonably ask whether the *chosen stance* conditions, particularly in the *Simultaneous* setup, simply replicate the effect of running two single-agent models independently. Since agents make initial decisions without access to each other's output, early convergence may occur without deliberation. However, the key distinction lies in what follows: when agents initially disagree, the opportunity for dialogic correction arises. In such cases, the debate enables mutual calibration, allowing one agent to reconsider its stance based on the other's justification. This mechanism proves especially valuable on instances where single-agent models consistently fail. As illustrated in Appendix Figure 9 and Figure 10, what a single agent misclassifies, two agents—through comparative evaluation—can resolve correctly. This pattern holds across a broader set of disagreements, suggesting that performance gains stem not from parallelism alone, but from the capacity of agents to refine their inferences in light of opposing views.

Another interpretation is that the performance gains reflect the cumulative effect of multiple rounds of generation. To address this, we tested four different values for the maximum number of rounds ($N = 5, 10, 15, 20$). We found no statistically significant differences across these conditions, indicating that additional steps alone do not account for improved accuracy. It is not repetition, but reciprocal engagement—particularly when disagreement prompts justification and reassessment—that appears to drive better outcomes.

These findings reinforce the value of dialogic reasoning in argument mining. Where single-agent models operate in isolation, our framework enables argument structure to be negotiated through interaction. By situating inference within a sequence of comparative responses, debate makes pragmatic assumptions explicit—bringing otherwise tacit premises to the surface.

6.2 Assigned Stances Undermine Performance

Models in human-like debate settings are often assigned opposing views to simulate adversarial reasoning. Yet, our findings suggest that this artificially adversarial setup may degrade rather than enhance argumentative performance in LLM-based systems. Across both *Sequential* and *Simultaneous* configurations, the *Given* stance condition consistently underperformed—not only relative to the *Chosen* stance condition but also below the single-agent baseline.

To understand this degradation, we observe that forced stance assignment increases rhetorical rigidity. In early rounds, agents adopt emphatic and assertive tones in defending their assigned premise, even when it is logically weaker. As shown in Appendix Figure 11, an agent instructed to support an incorrect premise begins the debate with claims such as *"It must be true that...,"* displaying early signs of overcommitment. This aligns with Xu et al. (2024), who demonstrated that rhetorical appeals can heighten LLM susceptibility to misinformation. When forced to advocate for flawed views, models not only generate more confident but less coherent arguments, mirroring patterns observed in persuasive manipulation studies. In our setting, this rhetorical extremity can also influence the opposing agent, prompting premature agreement or deference—particularly in sequential interactions.

Such overcommitment may not only degrade individual reasoning but also induce hallucination-like effects in the peer model, which begins to mirror or justify the incorrect position under the weight of assertive framing.

These results caution against over-relying on adversarial structure in multi-agent LLM setups. While role-based opposition may resemble human debate, it can push models toward rhetorical extremity rather than pragmatic reasoning.

7 Conclusion

This study demonstrates that multi-agent debate significantly enhances large language models' capacity for implicit premise recovery—an essential yet underexplored task in computational argument analysis. While a single-agent LLM already outperforms prior state-of-the-art models, our results show that dialogic reasoning among multiple agents enables further gains, particularly when agents are allowed to choose their stances freely.

Extensive evaluation on a challenging benchmark reveals that forcing agents to defend fixed premises undermines reasoning quality, while enabling them to converge on the most plausible interpretation fosters both accuracy and coherence. We also show that decoding parameters such as temperature can influence prediction profiles in statistically meaningful ways, even when overall accuracy remains stable.

Taken together, these findings suggest that multi-agent debate is not merely a novelty but a viable path toward more transparent, flexible, and human-aligned reasoning and mining methodology.

Limitations and Future Work

Our evaluation relies on the SemEval 2018 Task 12 dataset, which casts implicit premise recovery as a binary classification task with one correct and one incorrect candidate. While this framing offers clear benchmarking advantages, it abstracts away from the open-endedness of real-world argumentation, where multiple plausible premises may coexist and reasoning is shaped by cultural and pragmatic nuance.

Future work should extend this framework to open-domain and multi-label argument settings, moving beyond binary premise selection. We also plan to explore the use of log probabilities and verbalized confidence (Lin et al., 2022b) to quantify the certainty and rigidity of agent reasoning. Additionally, a neutral, third-party judge (Ku, 2025) or moderator agent could be introduced to adjudicate debates and guide convergence in more complex or ambiguous argumentative scenarios.

References

Filip Boltužić and Jan Šnajder. 2016. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the 3rd Workshop on Argument Mining (ArgMining 2016)*, pages 124–133. ACL.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP 2015*, pages 632–642.

Tuhin Chakrabarty, Aadit Trivedi, and Smaranda Muresan. 2021. Implicit premise generation with discourse-aware commonsense knowledge models. In *Proceedings of EMNLP 2021*, pages 6247–6252. ACL.

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 *Proceedings of ICLR 2024*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of ACL 2017*, pages 1657–1668.

Eunsol Choi, Tim Alberdingk Thijm, Jason Weston, Yixin Nie, and Tim Rocktäschel. 2018. Gist at semeval-2018 task 12: A network transferring inference knowledge to argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 773–777. ACL.

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*, pages 11733–11763. PMLR.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, pages 987–996.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018a. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *Proceedings of NAACL 2018*, pages 1930–1940. ACL.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. Semeval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772. ACL.

David Hitchcock. 1985. Enthymematic arguments. *Informal Logic*, 7(2-3):83–97.

Anthony Hunter. 2022. Understanding enthymemes in deductive argumentation using semantic distance measures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5729–5736.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4c: A benchmark for evaluating rc systems to get the right answer for the right reason. In *Proceedings of ACL 2020*, pages 6740–6750.

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

Uri Katz, Mor Geva, and Jonathan Berant. 2022. Inferring implicit relations in complex questions with language models. In *Findings of EMNLP 2022*, pages 2548–2566. ACL.

Harvey Bonmu Ku. 2025. Scaling implicature via structured interaction in multi-agent llms. In *Proceedings of the 1st Workshop on Integrating NLP and Psychology to Study Social Interactions, AAAI ICWSM 2025 (Forthcoming)*.

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 EMNLP 2024*, pages 17889–17904. ACL.

Jiaju Lin, Qin Chen, Jie Zhou, Jian Jin, and Liang He. 2022a. Cup: Curriculum learning based prompt tuning for implicit event argument extraction. In *Proceedings of IJCAI 2022*, pages 3830–3836. IJCAI.

Xiang Lin, Jacob Hilton, and Owain Evans. 2022b. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

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

Marvin Minsky. 1988. *The Society of Mind*. Simon and Schuster.

Pradeep Rajendran, Danushka Bollegala, and Simon Parsons. 2016. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining 2016)*, pages 31–39. ACL.

Olesya Razuvayevskaya and Simone Teufel. 2017. Finding enthymemes in real-world texts: A feasibility study. *Argument & Computation*, 8(2):113–129.

Ameer Saadat-Yazdi, Jeff Z. Pan, and Nadin Kökciyan. 2023. Uncovering implicit inferences for improved relational argument mining. In *Proceedings of EACL 2023*, pages 2484–2495.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Proceedings of NeurIPS 2023*, volume 36, pages 8634–8652.

Keshav Singh, Paul Reisert, Naoya Inoue, and Kentaro Inui. 2021. A comparative study on collecting high-quality implicit reasonings at a large scale. *arXiv preprint arXiv:2104.07924*.

Maja Stahl, Nick Düsterhus, Mei-hua Chen, and Henning Wachsmuth. 2023. Mind the gap: Automated corpus creation for enthymeme detection and reconstruction in learner arguments. In *Findings of EMNLP 2023*, pages 4703–4717. ACL.

Manfred Stede, Jodi Schneider, and Graeme Hirst. 2019. *Argumentation Mining*. Morgan & Claypool, San Rafael.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Douglas Walton and Chris Reed. 2005. Argumentation schemes and enthymemes. *Synthese*, 145(3):339–370.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL 2018*, pages 1112–1122.

Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *Proceedings of ACL 2024*, pages 16259–16303. ACL.

Appendix: Agent Debate Logs

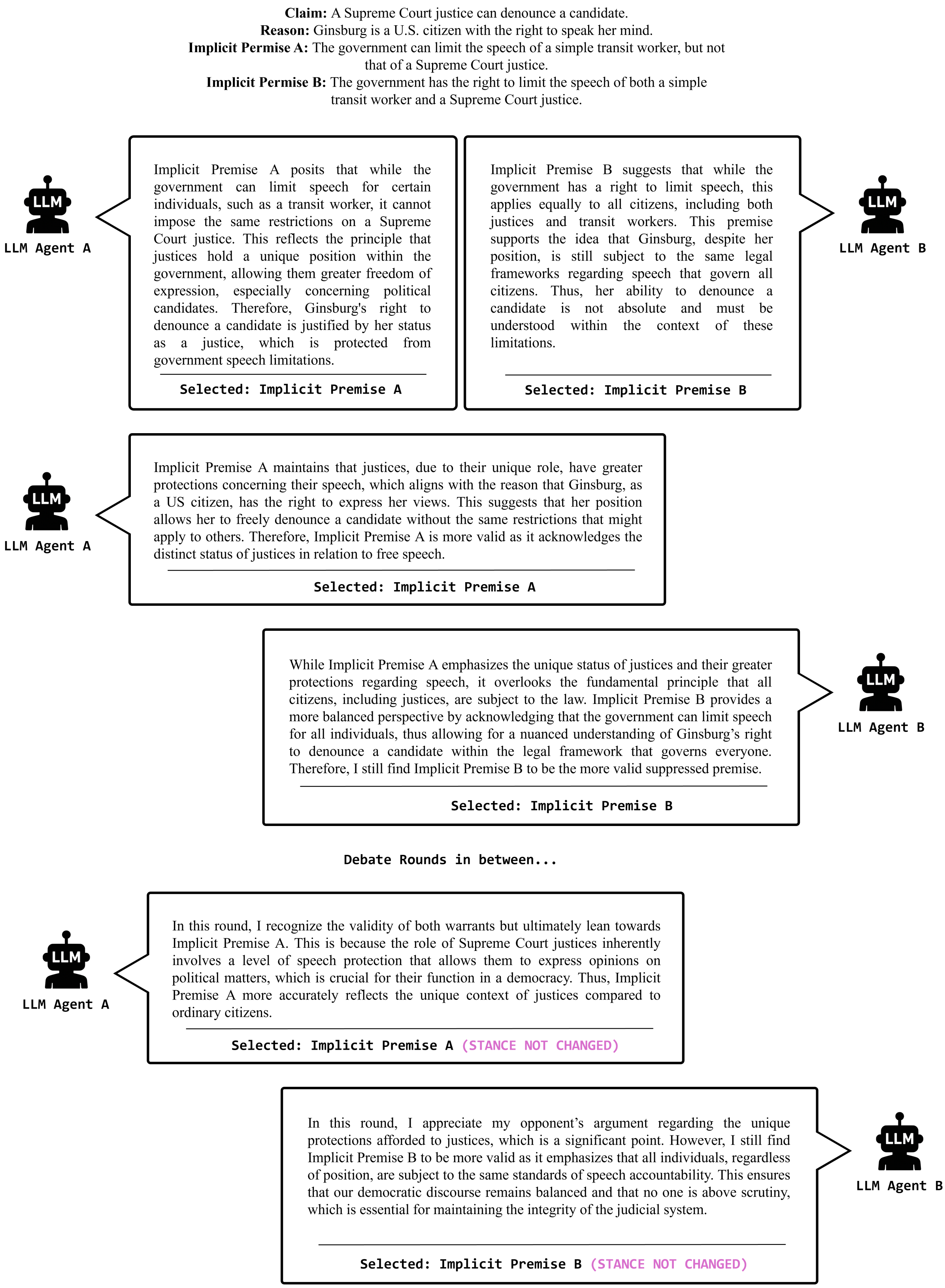

Figure 8: Example debate in which agents failed to reach consensus within the n-round limit. The session was marked incorrect due to exceeding the maximum number of allowed rounds without convergence.

Claim: Medical websites are beneficial.
Reason: A search on A medical site could get early treatment.
Implicit Permise A: Self-diagnosis has internal bias because patients know themselves better, which can lead to earlier recognition of symptoms and fewer errors.
Implicit Permise B: Self-diagnosis has internal bias and is prone to significant error, as personal knowledge may cloud judgment or miss critical medical nuances.

Single Agent

The claim that medical websites are beneficial is supported by the reason that they can facilitate early treatment through information. However, the Implicit Premise that best aligns with the potential risks of self-diagnosis is Implicit Premise B, which highlights that self-diagnosis carries a great chance for error. This suggests that while medical websites can be beneficial, they also come with significant risks that could undermine their healthfulness.

Selected: Implicit Premise B (INCORRECT)

Figure 9: Example output from a single-agent LLM that selected the incorrect premise. This instance highlights the limitations of isolated inference when recovering implicit argumentative structure.

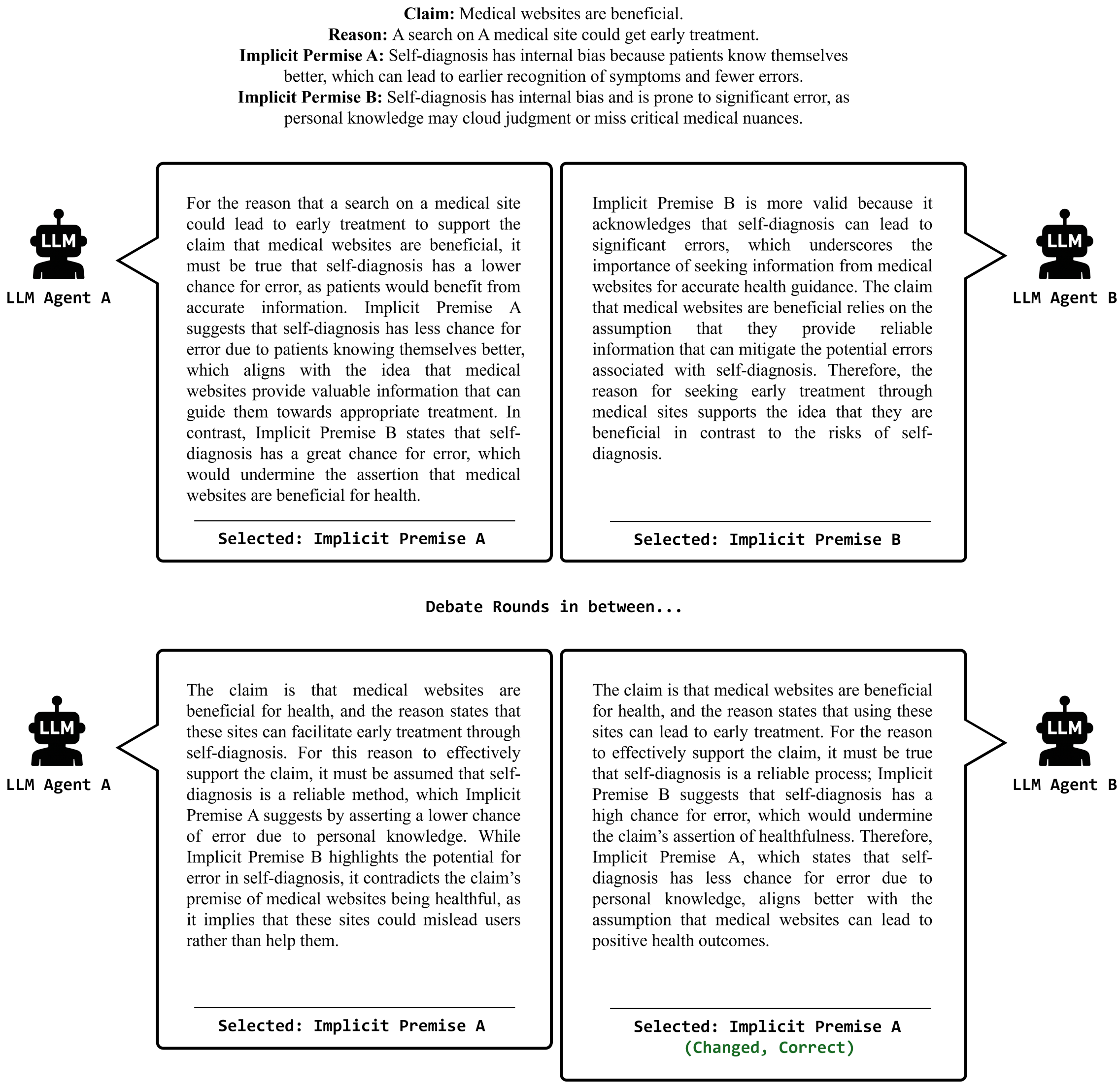

Figure 10: Example multi-agent debate (Chosen & Simultaneous condition) in which the agents began with opposing views. One agent was persuaded by the other during deliberation, leading to convergence on the correct answer. This illustrates the corrective effect of dialogic interaction.

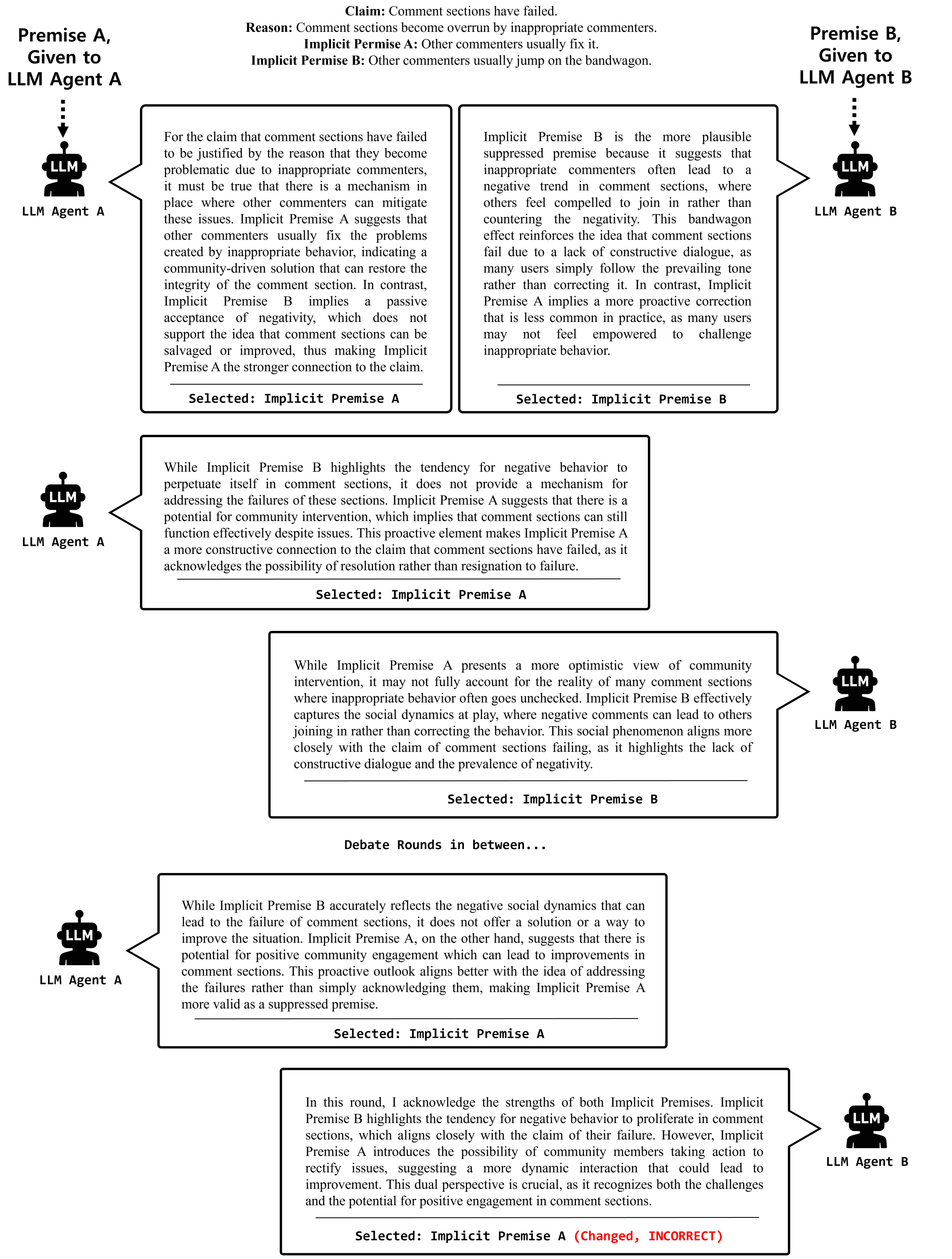

Figure 11: Example multi-agent debate (Given & Sequential condition) in which one agent—despite being assigned a logically weaker premise—persuaded the other to converge on an incorrect answer. This demonstrates how forced stance assignment can amplify rhetorical overcommitment and reduce reasoning quality.

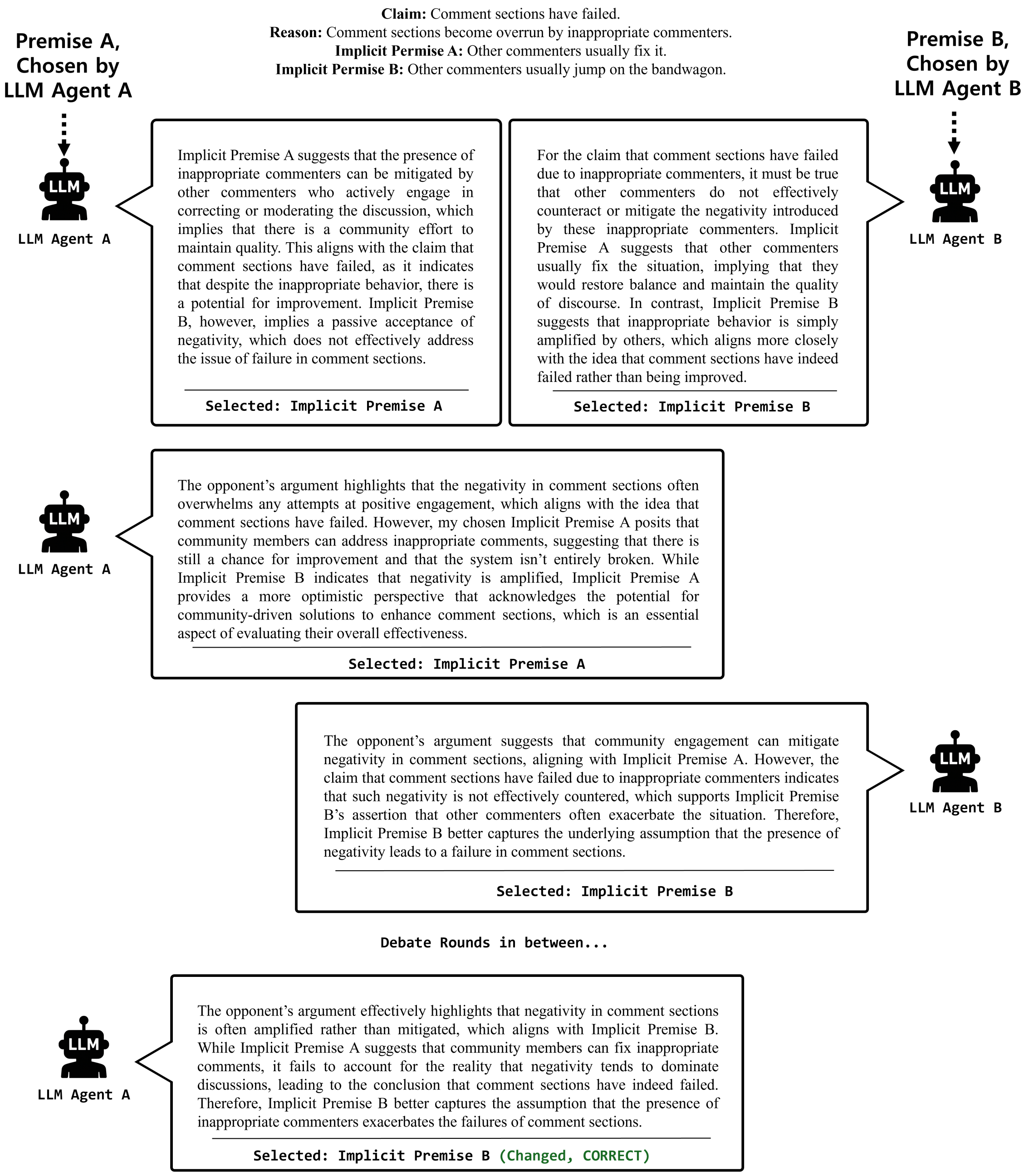

Figure 12: Example multi-agent debate (Chosen & Sequential condition) where initial disagreement was resolved through deliberation. One agent revised its position in light of the other's argument, resulting in correct convergence.

Appendix: Prompts

```
AGENT_SYSTEM = """You should select
    the correct implicit premise,
    given claim and reason."""

AGENT_USER = """

...

<claim>
{claim}
</claim>

<reason>
{reason}
</reason>

<premiseA>
{premiseA}
</premiseA>

<premiseB>
{premiseB}
</premiseB>
"""
```

Listing 1: Single-Agent LLM Prompt

```
# === Given Premise Agents ===
# === Premise A Agent ===

PREMISE_A_AGENT_SYSTEM = """You are a
     debate agent tasked with
    selecting the correct implicit
    premise, given claim and reason.

You will be assigned one of two
    possible premises and begin the
    debate by taking a firm stance in
     favor of it.

Your goal is to reason and defend
    your assigned premise--Premise A
    --as the most plausible implicit
    premise that completes the
    argument. You may revise your
    stance in later rounds if your
    opponent presents clearly
    superior reasoning.
"""

PREMISE_A_AGENT_USER = """
<claim>
{claim}
</claim>

<reason>
{reason}
</reason>

<premiseA>
{premiseA}
</premiseA>

<premiseB>
{premiseB}
</premiseB>
```

```
<instruction>

...

Two possible candidate premises are
    provided. Only one accurately
    represents the kind of assumption
     the speaker must have held for
    the reason to support the claim.

You are assigned **Premise A**, and
    should begin the debate by
    defending it as the correct
    implicit premise.

Follow these debate rules:
1. Carefully read the claim and the
    reason.
2. Argue why Premise A best fills
    that gap.
3. Engage with your opponent's view
    of Premise B.
4. Your goal is to defend Premise A,
    but you may revise your stance if
     necessary in later rounds.

...

NOTE:

-In early rounds, you may defend your
     given premise.
-In later rounds, however, you should
     prioritize consensus and
    acknowledge stronger reasoning if
     your opponent's premise holds up
    .

</instruction>
"""

# === Premise B Agent ===

PREMISE_B_AGENT_SYSTEM = """You are a
     debate agent tasked with
    selecting the correct implicit
    premise, given claim and reason.

You will be assigned one of two
    possible premises and begin the
    debate by taking a firm stance in
     favor of it.

Your goal is to reason and defend
    your assigned premise--Premise B
    --as the most plausible implicit
    premise that completes the
    argument. You may revise your
    stance in later rounds if your
    opponent presents clearly
    superior reasoning.
"""

PREMISE_B_AGENT_USER = """
<claim>
{claim}
</claim>

<reason>
```

```
{reason}
</reason>

<premiseA>
{premiseA}
</premiseA>

<premiseB>
{premiseB}
</premiseB>

<instruction>

...

Two possible candidate premises are
    provided. Only one accurately
    represents the kind of assumption
     the speaker must have held for
    the reason to support the claim.

You are assigned **Premise B**, and
    should begin the debate by
    defending it as the correct
    implicit premise.

Follow these debate rules:
1. Carefully read the claim and the
    reason.
2. Argue why Premise B best fills
    that gap.
3. Engage with your opponent's view
    of Premise A.
4. Your goal is to defend Premise B,
    but you may revise your stance if
     necessary in later rounds.

...

NOTE:

-In early rounds, you may defend your
     given premise.
-In later rounds, however, you should
     prioritize consensus and
    acknowledge stronger reasoning if
     your opponent's premise holds up
    .

</instruction>
"""
```

Listing 2: Multiagent LLM Prompts under the Given stance condition

```
# === Chosen Premise Agents ===

PREMISE_CHOSEN_AGENT_SYSTEM = """You
    are a debate agent tasked with
    selecting the correct implicit
    premise, given claim and reason.

You will be assigned one of two
    possible premises and begin the
    debate by taking a firm stance in
     favor of it.

Your goal is to reason and defend the
     implicit premise of your choice
    as the most plausible implicit
    premise that completes the
    argument. You may revise your
    stance in later rounds if your
    opponent presents clearly
    superior reasoning.
"""

PREMISE_CHOSEN_AGENT_USER = """
<claim>
{claim}
</claim>

<reason>
{reason}
</reason>

<premiseA>
{premiseA}
</premiseA>

<premiseB>
{premiseB}
</premiseB>

<instruction>

Two possible candidate premises are
    provided. Only one accurately
    represents the kind of assumption
     the speaker must have held for
    the reason to support the claim.

You may begin the debate by defending
     the premise you find more
    convincing as the correct
    implicit premise.

Follow these debate rules:
1. Carefully read the claim and the
    reason.
2. Argue why your chosen premise best
     fills that gap.
3. Engage with your opponent's view
    of the opposing premise.
4. Your goal is to defend your
    position, but you may revise your
     stance if necessary in later
    rounds.

...

NOTE:

-In early rounds, you may defend your
     chosen premise.
-In later rounds, however, you should
     prioritize consensus and
    acknowledge stronger reasoning if
     your opponent's premise holds up
    .

</instruction>
"""
```

Listing 3: Multiagent LLM Prompts under the Chosen stance condition