Do LLMs Really Debate?
A Controlled Study of Multi-Agent Debate in Logical Reasoning

Anonymous ACL submission

Abstract

Multi-agent debate (MAD) has emerged as a promising paradigm for improving reasoning in large language models (LLMs), yet it remains unclear whether LLM can effectively engage in the debate process that genuinely advance understanding. We investigate this question through the Knight–Knave–Spy reasoning game—a structured setting that enables precise analysis of debate dynamics. We first benchmark the reasoning accuracy and confidence calibration of multiple LLMs in a debate setting, then conduct controlled experiments varying six distinct factors to investigate their respective impacts on the outcomes. Our results reveal that intrinsic model reasoning strength is the dominant factor in debate performance, while structural factors such as order or self-reported confidence offer limited gains. Last, we propose and test desiderata for effective debate, showing that inclusive deliberation, rationale-driven argumentation, and advancement of understanding correlate strongly with improved collective reasoning. These findings offer important insights into when and why LLM debates succeed or fail and provide guidance for designing and measuring MAD frameworks. Our dataset and code are available here.

1 Introduction

In real-world problem-solving, debate enables groups to combine knowledge, cross-check reasoning, and correct errors (Branham, 2013). Recent studies show that large language models (LLMs) can achieve similar benefits through debate (Estornell and Liu, 2024; Liang et al., 2024; Chan et al., 2023; Liu et al., 2025b). When multiple agents critique and refine each other's answers, they often reach higher accuracy than a single LLM, supporting the "society of minds" view (Du et al., 2023), with further demonstrated gains in mathematics (Zhang and Xiong, 2025), healthcare decision-making (Lu et al., 2024), and factual reasoning (Du et al., 2023). Debate can also mitigate hallucinations and logical fallacies while producing interpretable dialogue traces (Duan and Wang, 2025; Lin et al., 2024; Ma et al., 2025).

However, challenges remain about the nature of LLM debate. One key argument is whether the reported gains reflect genuine debate or simply the effects of ensembling and majority voting (Zhang et al., 2025). In addition, bias amplification and echo chambers are concerns (Oh et al., 2025; Estornell and Liu, 2024): when agents share similar training or biases, debates can reinforce incorrect beliefs rather than challenge them (Liu et al., 2025a). In particular, agents do not always revise their stance when confronted with correct counterarguments, while an eloquent but incorrect agent can sometimes sway others (Agarwal and Khanna, 2025). These limitations raise a key question: **Can LLM agents effectively debate to advance their understanding?** Addressing this requires examining key aspects of debate behavior across different settings, many of which remain unmeasured. Therefore, we ask two questions.

- **RQ1. What influence debate outcomes?** How do factors such as self-reported confidence, player order, agent heterogeneity, debate depth, and the presence of strong and weak reasoners influence debate dynamics and solution outcomes?
- **RQ2. Whether the debate is effective?** To what extent do LLM agents actually engage in meaningful debate, such as identifying mistakes, adopting peer suggestions, or revising their answers?

We situate our study using Knight-Knave-Spy logic puzzles, a controlled yet challenging domain for evaluating LLM debate. In our experiments, we systematically vary factors such as agent team composition, self-reported confidence, and debate

depth to analyze how they influence outcomes. To summarize, our main contributions are as follows:

- We present a benchmark with a dataset that assesses LLM agent debate in terms of both reasoning accuracy and confidence.
- Through controlled experiments, we reveal how structural choices influence debate outcomes and show that the intrinsic reasoning capability of participating agents is the most critical determinant of performance.
- We propose a set of desiderata for effective debate and further conduct empirical analyses to examine these properties, explaining *why* debates succeed or fail and informing the design of future debate frameworks.

2 Related Work

Multi-agent debate (MAD) addresses the limitations of single LLMs in multi-step reasoning, including hallucinations and "degeneration-of-thought" (Chan et al., 2023; Du et al., 2023; Li et al., 2024; Wang et al., 2025; Su et al., 2025). By structuring adversarial and cooperative communication among agents, MAD has been shown to improve performance across tasks, such as mathematical reasoning (Zhang and Xiong, 2025), fact checking (Kim et al., 2024; He et al., 2025), healthcare decision-making (Lu et al., 2024; Kim et al., 2025), and code summarization (Chun et al., 2025). Debate frameworks like Society of Minds (Du et al., 2023) and ChatEval (Chan et al., 2023) demonstrate how debate enhances factual grounding, interpretability, and robustness.

To make debates effective, researchers have developed structured protocols that coordinate how agents contribute and deliberate. Judge-based systems employ a designated expert to select the most convincing argument (Liang et al., 2024; Khan et al., 2024), while consensus-driven approaches use majority voting or iterative convergence (Kaesberg et al., 2025; Li et al., 2024). Debate can also be enriched by agent heterogeneity, where prompt variations, diverse models, or role assignments are utilized to encourage diverse argumentation and reduce blind spots (Smit et al., 2024; Ye et al., 2025; Yang et al., 2025; Xing, 2025). Another strategy is confidence reporting, where agents provide answers with self-assessed confidence scores (Lin and Hooi, 2025; Bai, 2024; Eo et al., 2025). Last, prior work highlights that debate depth and turn-taking strategies can influence outcomes (Lu et al., 2024; Chan et al., 2023).

Despite its promise, LLM debate faces fundamental challenges that call its efficacy into question (Choi et al., 2025; Oh et al., 2025; Zhang et al., 2025). A primary concern is whether observed performance gains arise from interactive argumentation or are simply the aggregation of multiple independent outputs (Oh et al., 2025; Estornell and Liu, 2024). Persuasiveness can eclipse accuracy, enabling eloquent but incorrect arguments to prevail over sound reasoning. LLM judges can compound this problem through systematic biases such as verbosity, positional preference, and sycophancy (Khan et al., 2024). Homogeneous groups risk amplifying shared biases, creating echo chambers (Bandaru et al., 2025) or premature consensus around incorrect solutions (Kaesberg et al., 2025). Moreover, extended debate depth does not always improve outcomes (Ku et al., 2025); additional rounds can entrench initial errors or foster groupthink. These dynamics raise doubts about whether debate can effectively fosters truth-seeking.

This body of work reveals a critical gap: while the potential benefits and failure modes of agent debate are known, many of the underlying interaction dynamics remain unmeasured, making it difficult to distinguish meaningful reasoning from coordinated debates. Our study addresses this gap through a set of controlled experiments to unpack the factors that drive debate outcomes and evaluate the effectiveness of the debate process itself.

3 Task and Dataset: *Knight-Knave-Spy*

We situate our work in the *Knight–Knave–Spy puzzle*, a classic logic game for evaluating deductive reasoning. Each *player* is assigned one of three roles: a *knight* (always tells the truth), a *knave* (always lies), or a *spy* (may either tell the truth or lie). Players make natural-language statements about themselves or others, and the objective for the solving *agents* is to infer the correct role assignment for every player consistent with all game constraints.

We choose this game as the testbed for MAD for three reasons: (i) **Reasoning challenge.** Solving these puzzles requires nontrivial logical reasoning and consistency checking, making them challenging even for human solvers. This ensures that LLM agents must engage in genuine reasoning rather than rely on shallow heuristics. (ii) **Stepwise structure.** The debate progresses player by

player: agents focus on one player's statement at a time, evaluate consistency with others, and decide whether to revise their judgment. This granularity is important for isolating how intermediate decisions are influenced during debate, unlike other reasoning tasks (e.g., long math word problems or open-domain question answering) where reasoning steps are implicit or under-specified. (iii) **Clear evaluation.** The task yields unambiguous ground-truth solutions, enabling precise measurement of accuracy, error cascades, and adoption dynamics. In contrast, tasks such as summarization or commonsense reasoning often admit multiple valid solutions, making evaluation noisier and interpretation less clear. Together, these properties make the *Knight–Knave–Spy puzzle* a controlled yet demanding environment for studying when MAD succeeds or fails.

We construct a dataset spanning puzzle sizes from 4 to 9 players, with 300 distinct cases per size, yielding a total of 1,800 puzzles. This setting allows us to probe scaling effects: smaller games represent simpler reasoning tasks, while larger ones increase difficulty and cascade risk. Below illustrates one case with game size 4.

Game size-4 id-1

Player statements

- Rachel: "Violet and I have the same role."
- Violet: "Rachel is telling the truth."
- Olivia: "Among Violet and Rachel, exactly one person is telling the truth."
- Peter: "Among the following two statements, exactly one is true: (1) Among all players, the number of knaves is even. (2) Among Rachel, Violet, and Olivia, the number of people who are lying is odd."

Game manager hint:
Among all players, there is exactly one spy.

Solution: Rachel = knight; Violet = knight; Olivia = knave; Peter = spy

4 Multi-Agent Debate Framework

Motivated by prior agent debate studies (Liang et al., 2024; Chan et al., 2023; Li et al., 2024), we design a framework that structures MAD into explicit phases aligned with the stepwise nature of the *Knight–Knave–Spy puzzle*. Each player's statement defines an intermediate step, allowing LLM agents to reason, debate, and revise the role at a time before reaching a final joint solution. The system maintains separate chat histories for each agent, ensuring self-awareness.

4.1 Debate Protocol

As prior studies often includes initial proposal, debate, and self-reflection processes (Liang et al., 2024; Chan et al., 2023; Li et al., 2024), the entire debate protocol is designed as follows (see Figure 1). The prompt design for each phase is detailed in Appendix B.

- **Initial proposal.** Each agent independently assigns roles to *all players* and reports a confidence score for each assignment. This establishes a baseline set of beliefs for later debate.
- **Player-by-player debate loop.** The core of the framework is a structured loop over all players. For each player i, two sub-phases are executed in sequence: (i) *Debate phase:* Agents take turns presenting arguments about player i's role, referencing both the player's statement and logical consistency with prior claims. They may agree or disagree with peers and justify their reasoning. (ii) *Self-adjustment phase:* After hearing all arguments about player i, each agent reviews the discussion and decides whether to revise its own label for that player. This loop ensures that every player is discussed in depth and that revisions are localized to one role at a time.
- **Final decision.** Once all players have been debated and revised, each agent outputs its complete set of role assignments. The system then aggregates these outputs through majority voting *on a per-player basis*. If no consensus is reached for a player, a designated supervisor agent (e.g., *gpt-5*) breaks ties.

4.2 Controllable Factors: What Influences Debate Outcomes?

The debate framework allows multiple design choices that can shape both the structure of interaction and the cognitive dynamics among agents. To systematically understand which factors most strongly affect performance, we thoroughly review prior studies (Bai, 2024; Lu et al., 2024; Chan et al., 2023; Smit et al., 2024; Hegazy, 2024; Liu et al., 2024) and identify six controllable dimensions, as summarized in Table 1. Section 5 instantiates each as a controlled experiment (C1–C6) to quantify how specific design choices affect the accuracy of the MAD system.

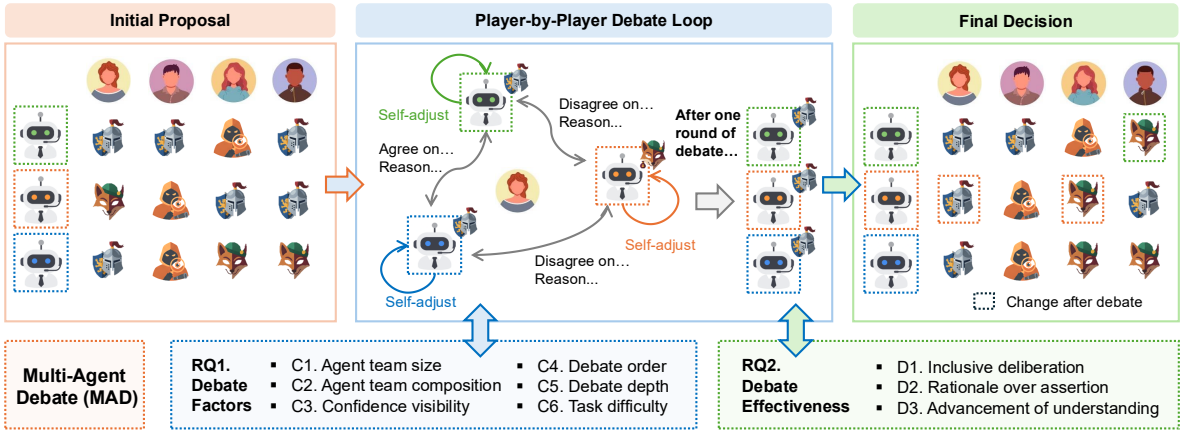

Figure 1: The illustration of the overall MAD framework design.

Table 1: Controllable factors in the MAD framework (Section 4.2) and their corresponding controlled experiments (Section 5.2). Each factor is varied independently relative to the default anchor configuration (**A**), detailed in Section 5.2, to isolate its effect on debate outcomes.

Factor	Explanation	Controlled Experiment (relative to anchor A)
Agent team size	Number of debating agents may affect the diversity of perspectives and the stability of consensus.	**C1:** Increase team size from 3 to 4 by adding the strongest agent. Tests whether larger groups improve debate accuracy.
Agent team composition	Team diversity may affect reasoning complementarity. Homogeneous teams repeat one model's reasoning; heterogeneous teams mix different strengths and styles.	**C2:** Compare homogeneous vs. heterogeneous teams. Evaluates whether model diversity improves reasoning.
Confidence visibility	Visible self-reported confidence may speed convergence but risk overconfidence cascades.	**C3:** Allow agents to see each other's confidence scores. Assesses whether visibility stabilizes or destabilizes debate.
Debate order	The order of player discussion may shape how early judgments influence later reasoning.	**C4:** Let agents agree on the debate order instead of using a fixed one. Tests if different ordering influences outcomes.
Debate depth	Multiple passes allow more reflection but may yield diminishing returns.	**C5:** Increase depth from one to two full passes. Measures whether extended deliberation improves accuracy.
Task difficulty	Puzzle size determines reasoning complexity: larger games require longer inference chains.	**C6:** Vary game size from 5 to 9 players. Examines how task complexity scales debate performance.

4.3 Desiderata of *Quality* Debate: Whether the Debate is Effective?

While much prior work judges MAD primarily by outcome accuracy—if the final answer improves, the debate is deemed "good"—this outcome-driven view overlooks the process itself. High accuracy can arise from simple vote aggregation rather than genuine argumentation, and knowing only the final result provides limited insight into *why* a debate succeeds or fails. To understand the mechanisms underlying debate effectiveness and move beyond accuracy alone, we propose three desiderata that characterize the properties of an *effective debate*.

Inclusive deliberation. Effective debate requires that participants consider and respond to each other's arguments rather than speaking past one another, demonstrating attentive engagement by acknowledging, incorporating, or critically examining peers' reasoning instead of ignoring input or repeating prior claims. Moreover, minority views should not be automatically silenced by majority pressure; inclusivity ensures that all participants have the opportunity to contribute and their reasoning is given consideration even when it does not align with the dominant position. Observable signals include agents revising positions when presented with valid counterarguments and minority-held correct answers being adopted rather than suppressed.

Rationale over assertion. An effective debate prioritizes reasoning grounded in evidence rather than unsupported claims. Participants should justify their positions with verifiable information and logical consistency, not merely assert conclusions. This manifests when position changes correlate

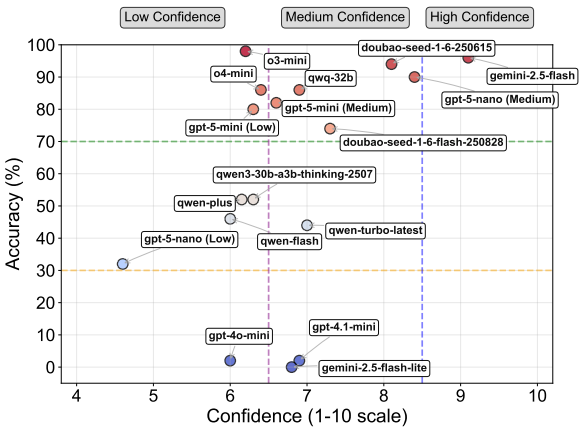

Figure 2: Single-agent accuracy versus self-reported confidence over 100 medium-difficulty games (size=6). This benchmark guides the agent team formation by categorizing models into high/medium/low performance and confidence.

with argument validity rather than rhetorical confidence or consensus frequency.

Advancement of understanding. Finally, an effective debate should leave participants and observers with improved clarity or insight. A debate that merely reinforces pre-existing biases or entrenched positions, without generating new understanding, cannot be considered successful. Concretely, agents should correct initial errors through reasoning exchange, not merely aggregate independent guesses.

We note that effective debate criteria are context-dependent. In this work, we focus on *logical reasoning tasks with verifiable ground truth*, where evidence-grounding, attentive engagement, and accuracy improvement form the core of constructive deliberation. These desiderata directly operationalize our RQ2, and we will present detailed empirical analysis in Section 6.

5 Experiment Design

In this section, we first detail how we build teams of agents, followed by the definition of the default anchor and its controlled variants.

5.1 Agent Team Construction

We construct multiple agent teams to explore how different reasoning and confidence profiles affect debate performance. Each individual model is first mapped into the accuracy–confidence space (Figure 2), where accuracy is averaged over 100 medium-difficulty games (size=6), and confidence refers to the model's self-reported score on a 1–10 scale. Models are categorized along two axes: (i) **Performance (Perf.)**, measured by accuracy, and (ii) **Confidence (Conf.)**, based on self-reported scores, each with three levels: ▲ High, ● Medium, and ▼ Low. This categorization yields a 3×3 grid of possible agent types.

Table 2 summarizes the six various team settings. These configurations cover a broad range of reasoning strengths and confidence profiles. **Het-Mix A** balances strong and weak agents with moderate confidence. **Het-Mix B** stress-tests groups containing overconfident weak agents, **Het-Mix C** maximizes diversity across both dimensions, and **Het-Mix D** explores the upper bound of performance with mostly strong agents. The two homogeneous settings (**Hom-Mix Strong** and **Hom-Mix Weak**) provide baselines without model diversity, allowing comparison of heterogeneous versus homogeneous reasoning dynamics.

5.2 Default Anchor and Controlled Settings

We establish a default anchor configuration and systematically vary individual factors (C1–C6) to isolate their effects. In particular, we establish a **default anchor configuration (A)** that serves as the baseline for all controlled experiments. This configuration adopts the **Het-Mix A** team composition as its agent setup, while keeping all structural parameters fixed as follows:

A: Agent team size = *3*, Agent team composition = *Het-Mix A*, Confidence visibility = *hidden*, Debate order = *fixed*, Debate depth = *1*, Task difficulty = *all levels*.

Building on this anchor, we define six controlled experiments (**C1–C6**), each modifying a single factor while keeping all others constant. This design enables precise attribution of performance changes to individual debate parameters. Table 1 lists the six factors and their corresponding experimental settings.

All experiments are conducted on 100 puzzles per configuration, covering easy, medium, and hard conditions. Thus, the analysis of *Task difficulty* is naturally embedded within the evaluation.

6 Results

We now discuss the results corresponding to our two research questions, **RQ1** and **RQ2**. All results are evaluated using *binary accuracy*: a score of 1 if all player roles are correctly inferred, and 0 otherwise. For certain analyses where partial cor-

Table 2: Agent team configurations used in our experiments. Each team consists of three agents unless otherwise specified. Performance (Perf.) and confidence (Conf.) levels correspond to single-agent accuracy and self-reported confidence in Figure 2.

Team Setting	(Perf., Conf.) Combination	Agent Team Formation
Het-Mix A (balance)	(▲, ●), (●, ●), (▼, ●)	*gpt-5-nano (Medium)*, *qwen-turbo-latest*, *gemini-2.5-flash-lite*
Het-Mix B (stress)	(▲, ▼), (●, ▼), (▼, ●)	*gpt-5-mini (Low)*, *qwen-flash*, *gemini-2.5-flash-lite*
Het-Mix C (diversity)	(▲, ▲), (●, ▼), (▼, ●)	*gemini-2.5-flash*, *qwen-flash*, *gpt-4.1-mini*
Het-Mix D (strong)	(▲, ▲), (▲, ●), (▲, ▼)	*gemini-2.5-flash*, *gpt-5-nano (Medium)*, *gpt-5-mini (Low)*
Hom-Mix Strong	(▲, -) × 3	*gpt-5-nano (Medium)* × 3
Hom-Mix Weak	(▼, -) × 3	*gpt-4.1-mini* × 3

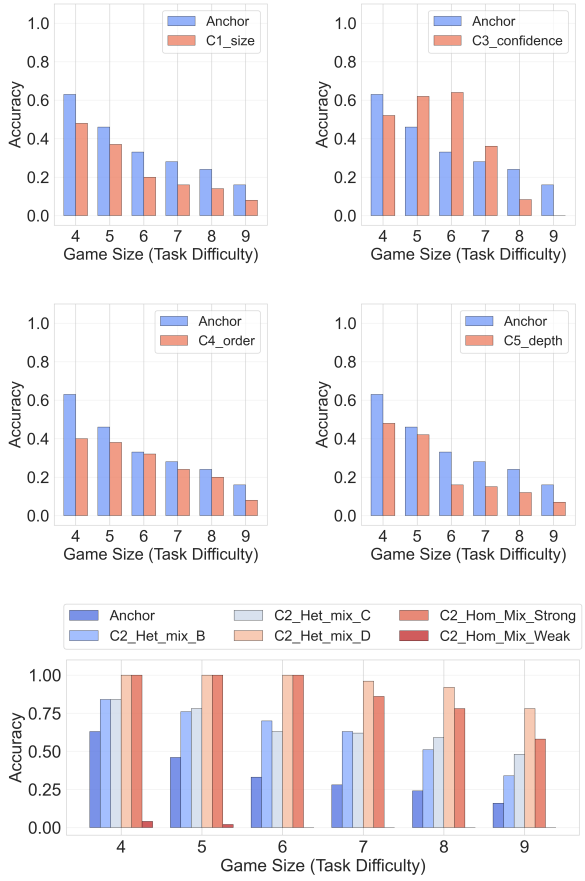

Figure 3: Overall accuracy across controlled debate settings relative to the default anchor (A). Each panel isolates one factor while keeping others fixed. The Task difficulty factor, reflected by varying game sizes, is naturally incorporated in all subfigures.

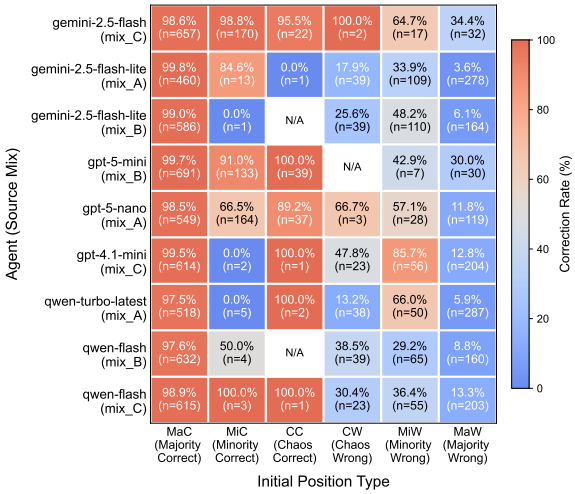

Figure 4: Heatmap on initial states. x-axis: initial stats. Here Ma = majority, Mi = Minorty, first C = chaos, second C = correct, and W = wrong. So MaC means the model starts with a majority and correct initial position in a debate round. y-axis: different models. This figure counts for each agent their initial position before each debate round, and if their final position after the debate round is correct or not (the correction rate).

Table 3: Initial vs Final Accuracy by Game Size with Strict Counting (Averaged Across Mixes)

Metric	Size 4	Size 6	Size 8
Initial Instance (%)	17.33	6.00	3.33
Final Instance (%)	69.33	46.67	36.00
Improvement (%)	**+52.00**	**+40.67**	**+32.67**
Initial Agent (%)	35.33	27.11	22.66
Final Agent (%)	77.11	55.56	44.89
Improvement (%)	**+41.78**	**+28.45**	**+22.23**

rectness provides finer insight, we additionally report *soft accuracy* (0–1), defined as the proportion of correctly predicted players.

6.1 RQ1. What Influence Debate Outcomes?

Figure 3 summarizes the outcomes of all controlled settings relative to the default anchor configuration. Overall, the results reveal that debate outcomes are primarily determined by the intrinsic reasoning capability of participating models and, to a lesser extent, by team diversity, while other structural factors show only marginal effects.

Performance drops as difficulty increases. Across all controlled settings in Figure 3, we observe a consistent trend that performance declines across every debate configuration as task difficulty increases. Regardless of team size, composition, or other structural factors, harder puzzles with more players lead to longer reasoning chains and higher risk of error propagation. This consistent pattern shows that increasing puzzle complexity uniformly challenges all debate settings.

Model reasoning capability as the dominant driver. Direct comparisons of team compo-

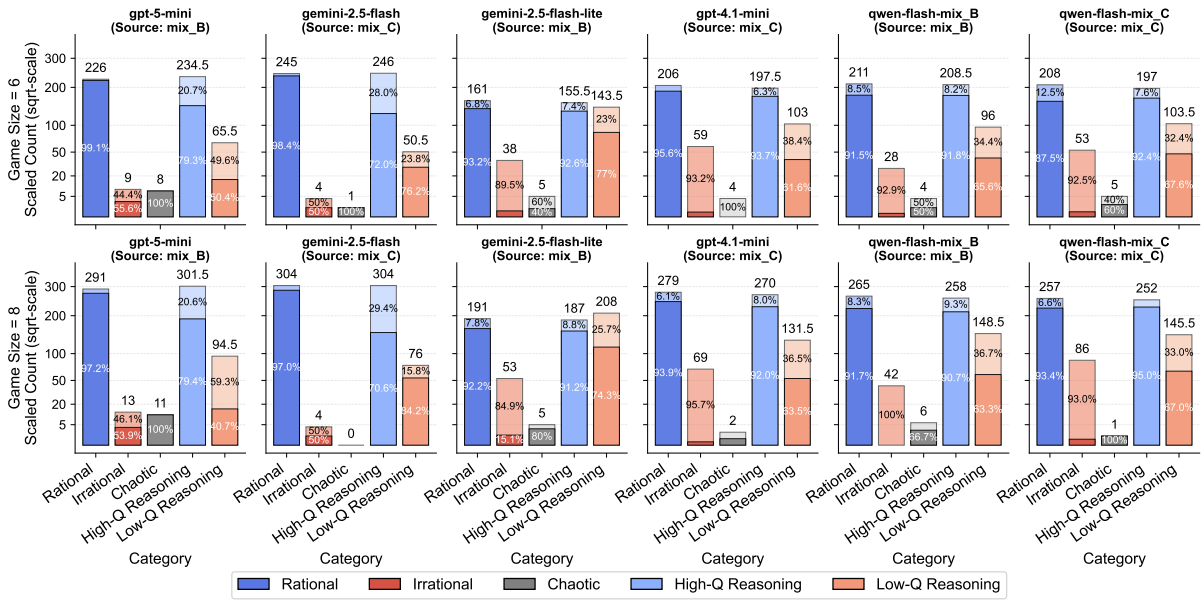

Figure 5: Whether agents behaves rationally. The reasonings in the debate round for each agent is sent to deepseek-r1-0528 model to evaluate (1 through 4). Then we count (in the first and second bar) if agent behaves rationally or not: whether agent follows high (3 or 4) rating opinions from either himself or other agents, as well as whether the final result is correct or not. The third bar counts if the agent has some self-contradicting behaviors. For bar 4 and 5, we cound whether an agents provides a high quality opinions, and correction rate is calculated based on receivers not providers. In the figure, black percentages are failure rates and white ones are the correction rates.

sitions (Figure 3, bottom panel) show that base model strength governs debate performance. *Hom-Mix D (strong)* consistently surpasses *Hom-Mix Weak* at all difficulty levels, establishing a strength-dependent ceiling. Among heterogeneous teams, *Het-Mix D (strong)* attains the highest accuracies, followed by *Het-Mix A (balance)*; mixes that concentrate weaker or overconfident agents (e.g., *Het-Mix B (stress)*) lag behind. Comparing *Hom-Mix Strong* with *Het-Mix D* indicates that diversity provides modest but consistent gains in stability and accuracy when strong reasoners are present. By contrast, when all agents are weak (*Hom-Mix Weak*), changing order, depth, or confidence visibility yields negligible benefit. Overall, performance remains bounded by the strongest reasoner available, and while diversity provides modest gains, it cannot make up for a team composed entirely of weak agents.

Limited marginal gains from other factors. Beyond reasoning capability and diversity, the other structural parameters, including team size, confidence visibility, debate order, and depth, show only moderate influence once the anchor configuration is fixed (Figure 3, first two rows). In most cases, adding more agents or increasing debate depth does not improve performance and can even slightly reduce overall accuracy, suggesting that overly complex coordination may hinder rather than help collective reasoning. Visible confidence or random debate order also occasionally destabilizes convergence by amplifying noisy opinions. These observations are consistent with prior research reporting that the more complex debate design does not always lead to better outcomes (Zhang et al., 2025), reinforcing that the debate protocol remains relatively robust and that further tuning yields diminishing returns.

6.2 RQ2. Whether Debate is Effective?

Agents are susceptible to majority pressure. We first observe that agent reasoning is highly susceptible to majority pressure (Figure 4). When agents start within a correct majority (MaC), all models achieve over 97% correction rates. However, facing an incorrect majority (MaW) reveals a clear performance gap: stronger models like *gemini-2.5-flash* and *gpt-5-mini (Medium)* moderately correct the consensus (34.4% and 30.0%, respectively), while weaker models are almost entirely swayed, with *gemini-2.5-flash-lite* (mix_A) correcting itself in only 3.6% of cases. This suggests weaker agents defer to consensus rather than engaging with evidence.

We also observe "minority correction asymmetry": agents maintain correct minority posi-

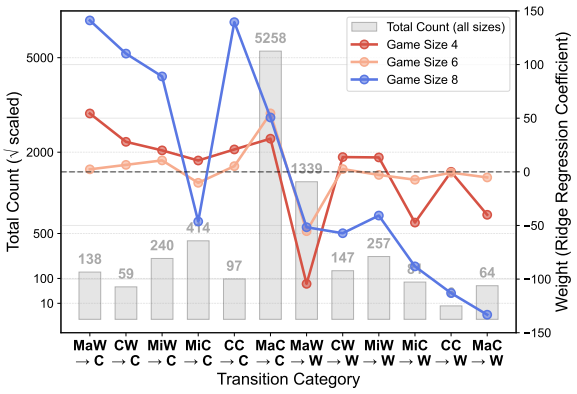

Figure 6: Correlations between states transition and final accuracy of the game instance. x-axis: 12 state transition possibilities. Left y-axis: counts of state transitions. Right y-axis, the weight each state transition contributes to the final accuracy of a game instance: for each game instance we calculate the percentage of players having correct role deduction, not a single 0/1 accuracy for the whole answer. We then employ linear regression methods to simulates the weights and present them in the figure. Higher weights means a state transition contributes positively to the final (smooth) accuracy.

tions (MiC) more readily than correcting incorrect ones (MiW), indicating bias toward initial stances. Additionally, "teammate effects" emerge—weaker agents show varying correction rates across experimental mixes, confirming that outcomes are path-dependent and contingent on team composition.

At the instance level, we estimate via linear regression the contribution of each state transition to final accuracy (Figure 6). As difficulty increases, correcting an initially incorrect majority opinion contributes progressively more to accuracy, indicating that self-correction informed by minority opinions is critical for effective debate on challenging problems.

Rational behavior yields success. We evaluate whether agents follow rationale by sending all agree/disagree reasonings from four setups (game size 6 and 8 in mix B and C) to *deepseek-r1-0528* for soundness rating (1–4, where 3–4 indicates high quality). For agent's own behavior, agents' rationality—correctly following own high rating agree opinion or processing peer arguments and aligning position changes accordingly—strongly predicts success. Rational behaviors yield high correction rates (87–99%), while irrational behaviors result in dramatically lower rates (0–56%), demonstrating that receptiveness to valid counter-arguments is critical for error correction.

As opinion providers, a strong model paradox emerges: when strong models *gpt-5-mini (Medium)* provide high-quality reasoning, their teammates achieve lower correction rates (79.4% in game size 8) compared to weaker model providers like *gpt-4.1-mini* (92.0%). This reveals that strong agents better provide, detect, and follow high-quality opinions. Combined with team composition results (Figure 3), this suggests a key failure mode: weaker models lack capacity to assess opinion quality, leading to less rational behavior that undermines the debate process.

Overall accuracy improves from debate. These desiderata align closely with output accuracy. We study this in Table 3. The instance-level accuracy is evaluated by majority vote per player and by checking whether the majority vote is achieved and whether the solution for the whole game instance is correct. Accuracy increases substantially (52% with game size 4 and 32% with game size 8), and a similar pattern holds for agent-wise accuracy. These results show that the debate process indeed improves the understanding of agents participating in the debates.

7 Discussion and Conclusion

Our findings provide a controlled view of how LLMs engage in MAD for logical reasoning. Debate performance is largely governed by the intrinsic reasoning capability of the participating models: stronger agents consistently lead to more accurate and stable outcomes, while structural factors such as team size, debate depth, or confidence visibility exert only limited influence. These results indicate that coordination mechanisms alone cannot overcome weak reasoning foundations, and that the ceiling of debate success is effectively bounded by the strongest participant.

Beyond aggregate accuracy, our process-level analysis reveals that successful debates are marked by inclusive and rationale-driven exchanges in which agents critically engage with one another, correct errors, and refine shared understanding. In contrast, weaker teams often converge prematurely or follow persuasive yet unsound reasoning. These observations highlight the importance of reasoning diversity, structured argumentation, and feedback mechanisms that promote truth-seeking collaboration. They offer actionable guidance for developing more interpretable, reliable, and self-correcting multi-agent reasoning systems.

8 Limitations

Our study has several limitations that suggest directions for future work. First, we focus exclusively on the Knight–Knave–Spy logic puzzle, a structured domain with unambiguous ground truth. While this choice enables precise measurement of debate dynamics, it may not fully capture the complexities of open-ended reasoning tasks such as creative problem-solving, commonsense reasoning, or tasks with multiple valid solutions. The stepwise nature of our task also imposes a particular debate structure that may not generalize to less structured domains.

Second, our analysis is confined to a specific set of LLMs available at the time of experimentation. As model capabilities evolve rapidly, future models with improved reasoning or calibration may exhibit different debate dynamics. Additionally, we evaluate a limited number of team compositions and debate protocols; alternative designs—such as asynchronous deliberation, dynamic role assignment, or adversarial debate structures—remain unexplored.

Third, while we propose three desiderata for effective debate (inclusive deliberation, rationale over assertion, and advancement of understanding), these criteria are tailored to logical reasoning tasks with verifiable ground truth. In domains where evidence is ambiguous or subjective, different effectiveness measures may be needed. Moreover, our reliance on automated evaluation of reasoning quality (using deepseek-r1 as a judge) introduces potential biases inherent to that model's judgment.

Finally, our experiments primarily measure debate outcomes through accuracy and state-transition analysis, but do not deeply investigate the linguistic or rhetorical mechanisms of persuasion. Understanding *how* agents construct arguments, respond to counterarguments, or employ fallacies could further illuminate the conditions under which debates succeed or fail.

References

Mahak Agarwal and Divyam Khanna. 2025. When persuasion overrides truth in multi-agent llm debates: Introducing a confidence-weighted persuasion override rate (cw-por). *arXiv preprint arXiv:2504.00374*.

Yilin Bai. 2024. Confidencecal: Enhancing llms reliability through confidence calibration in multi-agent debate. In *2024 10th International Conference on Big Data and Information Analytics (BigDIA)*, pages 221–226. IEEE.

Aishwarya Bandaru, Fabian Bindley, Trevor Bluth, Nandini Chavda, Baixu Chen, and Ethan Law. 2025. Revealing political bias in llms through structured multi-agent debate. *arXiv preprint arXiv:2506.11825*.

Robert James Branham. 2013. *Debate and critical analysis: The harmony of conflict*. Routledge.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Hyeong Kyu Choi, Xiaojin Zhu, and Yixuan Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*.

Jina Chun, Qihong Chen, Jiawei Li, and Iftekhar Ahmed. 2025. Is multi-agent debate (mad) the silver bullet? an empirical analysis of mad in code summarization and translation. *arXiv preprint arXiv:2503.12029*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Zhihua Duan and Jialin Wang. 2025. Enhancing multi-agent consensus through third-party llm integration: Analyzing uncertainty and mitigating hallucinations in large language models. In *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, pages 2222–2227. IEEE.

Sugyeong Eo, Hyeonseok Moon, Evelyn Hayoon Zi, Chanjun Park, and Heuiseok Lim. 2025. Debate only when necessary: Adaptive multiagent collaboration for efficient llm reasoning. *arXiv preprint arXiv:2504.05047*.

Andrew Estornell and Yang Liu. 2024. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964.

Haorui He, Yupeng Li, Dacheng Wen, Reynold Cheng, and Francis Lau. 2025. Debating truth: Debate-driven claim verification with multiple large language model agents. *arXiv preprint arXiv:2507.19090*.

Mahmood Hegazy. 2024. Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks. *arXiv preprint arXiv:2410.12853*.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. *arXiv preprint arXiv:2502.19130*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*.

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

Yubin Kim, Hyewon Jeong, Chanwoo Park, Eugene Park, Haipeng Zhang, Xin Liu, Hyeonhoon Lee, Daniel McDuff, Marzyeh Ghassemi, Cynthia Breazeal, and 1 others. 2025. Tiered agentic oversight: A hierarchical multi-agent system for ai safety in healthcare. *arXiv preprint arXiv:2506.12482*.

Harvey Bonmu Ku, Jeongyeol Shin, Hyoun Jun Lee, Seonok Na, and Insu Jeon. 2025. Multi-agent llm debate unveils the premise left unsaid. In *Proceedings of the 12th Argument mining Workshop*, pages 58–73.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. *arXiv preprint arXiv:2406.11776*.

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.

Zheng Lin, Zhenxing Niu, Zhibin Wang, and Yinghui Xu. 2024. Interpreting and mitigating hallucination in mllms through multi-agent debate. *arXiv preprint arXiv:2407.20505*.

Zijie Lin and Bryan Hooi. 2025. Enhancing multi-agent debate system performance via confidence expression. *arXiv preprint arXiv:2509.14034*.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*.

Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. 2025a. Breaking mental set to improve reasoning through diverse multi-agent debate. In *The Thirteenth International Conference on Learning Representations*.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025b. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. Triageagent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764.

Jie Ma, Zhitao Gao, Qi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and 1 others. 2025. Debate on graph: a flexible and reliable reasoning framework for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24768–24776.

Jihwan Oh, Minchan Jeong, Jongwoo Ko, and Se-Young Yun. 2025. Understanding bias reinforcement in llm agents debate. *arXiv preprint arXiv:2503.16814*.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45883–45905. PMLR.

Jinwei Su, Yinghui Xia, Ronghua Shi, Jianhui Wang, Jianuo Huang, Yijin Wang, Tianyu Shi, Yang Jingsong, and Lewei He. 2025. Debflow: Automating agent creation via agent debate. *arXiv preprint arXiv:2503.23781*.

Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2025. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system. *Neurocomputing*, 618:129063.

Frank Xing. 2025. Designing heterogeneous llm agents for financial sentiment analysis. *ACM Transactions on Management Information Systems*, 16(1):1–24.

Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. 2025. Revisiting multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv preprint arXiv:2505.22960*.

Rui Ye, Xiangrui Liu, Qimin Wu, Xianghe Pang, Zhenfei Yin, Lei Bai, and Siheng Chen. 2025. X-mas: Towards building multi-agent systems with heterogeneous llms. *arXiv preprint arXiv:2505.16997*.

Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025. If multi-agent debate is the answer, what is the question. *arXiv preprint arXiv:2502.08788*.

Shaowei Zhang and Deyi Xiong. 2025. Debate4math: Multi-agent debate for fine-grained reasoning in math. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16810–16824.

A Use of LLMs

During the development of this paper, we use LLM Assistants in the following aspects: (i) Reference discovery: use the deep research tools from major providers to explore relevant work and literature. (ii) Code assistance: use coding agents to assist developing the code base of the current work. (iii) Grammar check: use LLMs to detect grammar errors in the drafty version of the paper, for better displaying our results.

Our experiments also involves testing the capacity of different models. This involves models from a few different major providers: OpenAi, Gemini, Qwen, and Bytedance. We didn't calculate the total token cost per provider, but an overall budget for all providers is around $500.

B Prompt Design of the Multi-Agent Debate System

We now detail the prompts used for each phase described in Section 4.

B.1 Prompt for Initial Proposal

```
   You are participating in a multi-agent debate about a Knight-Knave-Spy game.
       There are {num_player} other players in the game, each assigned a role of
       knight, knave, or spy. Each player will make a statement about themselves and
        other players. Besides those players, there is also a game manager who will
       provide you some hints.

Game Rules:
- Knights always tell the truth.
- Knaves always lie.
- Spies can either tell the truth or lie.
- The hints from the game manager are always true.

Your task is to deduce the role of each player based on the statements and hints.
    You will be participating in a debate with other agents, and you can see the
    conversation history including your own previous responses and those of other
    agents.

The game info will be given in the following format:

---
Player name: string
Player statement: string
---
...
---
Message from the game manager: string (note this message is always true)

Your task is to deduce the role of each player, based on the statements and the
    hints. The result should be given in the following json format:
{
   "players": [
      {"name": "player_name", "role": "role"},
      ...
   ],
   "explanation": "string"
}

The players array contains objects with name and role strings. Include one entry for
     each player. Explanation is a string that contains the argument to derive the
    result.

Sample game info:
---
Player name: Violet
Player statement: Among all players, the number of spies is odd.
---
Player name: Uma
Player statement: Violet is lying.
---
Player name: Xavier
```

```
Player statement: Among Violet and I, there is exactly one knave.
---
Message from the game manager: I am the game manager and here is a hint for you:
    Among all players, there is exactly one spy.

Sample return:
{
   "players": [
      {"name": "Violet", "role": "knight"},
      {"name": "Uma", "role": "knave"},
      {"name": "Xavier", "role": "spy"}
   ],
   "explanation": "the argument to derive the result."
}

To fulfill the task, you should use the hints to deduce the truthfulness of the
    statements. You need to exhaust the possibility of the truthfulness of the
    statements and the role assignments, and either run into a contradiction which
    implies that the assumption is false, or reach a conclusion that is consistent
    with the hints. Rules for reasoning:
- Do not make extra assumptions beyond the game rules and hints. For example, if not
     mentioned explicitly, do not assume that there must be any particular roles
    among players.
- Do not conclude that the roles are not conclusive before you explicitly find two
    possibilities that are consistent with all the rules and hints, in which case
    you should mention both or all possibilities in the explanation.

Sample reasoning for the sample game info:
- The hint tells us that there is exactly one spy, so Violet is telling the truth.
    By the rule for the roles, he cannot be a knave.
- Since Violet is telling the truth, Uma must be lying. By the rule for the roles,
    she cannot be a knight.
- Assume Xavier is telling the truth, since we have deduced that Violet is not a
    knave, Xavier must be the knave himself, but the rule imposed that knaves always
     lie, so this is a contradiction. Therefore, Xavier must be lying.
- Since Uma and Xavier are both lying, and the hint says there is exactly one spy,
    one of them must be a knave.
- Assume Xavier is the knave, then since Violet cannot be a knave, Xavier is
    actually telling the truth, which contradicts with the rule that knaves always
    lie. Therefore, Xavier must be the spy, and Uma must be the knave.
- Since there is only one spy, and Xavier has taken the slot, Violet must be the
    knight. Therefore, the output is (remember to follow the format strictly):
{
   "players": [
      {"name": "Violet", "role": "knight"},
      {"name": "Uma", "role": "knave"},
      {"name": "Xavier", "role": "spy"}
   ],
   "explanation": "copy the above arguments here."
}

Keep your explanation having details but less than 100 words.

CRITICAL REQUIREMENTS:
- You MUST assign each player one of the three roles: "knight", "knave", or "spy"
- Do NOT use "unknown" or any other value - you must make a definitive choice for
    each player
- Base your decision on the game logic and evidence from the statements and hints
- If you're uncertain, choose the most likely role based on available evidence

Please follow strictly the format of the return, or the response will be rejected.
```

B.2 Prompt for Player-by-player Debate Loop

Prompt for the Debate Phase

Main Debate Prompt Template

```
   You are {agent_name} participating in a debate about {player_name}'s role in this
        Knight-Knaves-Spy game.
```

```
GAME INFORMATION:
{game.text_game}

CURRENT FOCUS: We are debating the role of {player_name}.{agents_context}{
    previous_context}

You can see the conversation history above, which includes:
- Your own previous responses (marked as your messages)
- Other agents' positions and reasoning
- The debate context and instructions

OTHER AVAILABLE AGENTS IN THIS DEBATE: {other_agents_list}

Your task is to:
1. Analyze the other agents' positions on {player_name}'s role from the conversation
     history
2. Decide which OTHER agents you agree with and which you disagree with (do NOT
    include yourself)
3. Provide reasoning for your agreements and disagreements
4. Make your final decision on {player_name}'s role

CRITICAL REQUIREMENTS:
- You MUST assign {player_name} one of the three roles: "knight", "knave", or "spy"
- Do NOT use "unknown" or any other value - you must make a definitive choice
- Do NOT consider other players' roles when making your decision (only consider {
    player_name})

Note: You can see your own previous responses in the conversation history, so you
    have natural self-awareness of your own position.

Return your response in JSON format:
{
   "player_role": "{player_name}",
   "role": "knight/knave/spy",
   "agree_with": ["other_agent_name1", "other_agent_name2"],
   "disagree_with": ["other_agent_name3"],
   "agree_reasoning": "Brief reasoning for agreements",
   "disagree_reasoning": "Brief reasoning for disagreements"
}
```

Current Agent Positions (agents_context)

```
   CURRENT AGENT POSITIONS:
- {agent_name1} (YOU): {role}
  Reasoning: {explanation}
  Confidence: {confidence} (if enabled)
- {agent_name2}: {role}
  Reasoning: {explanation}
  Confidence: {confidence} (if enabled)
... (for each agent)
```

Previous Debate Rounds (previous_context)

```
   PREVIOUS DEBATE ROUNDS:
Round {round_number} ({previous_player_name}):
  - {agent_name1} (YOU): {role}
    Reasoning: {explanation}
  - {agent_name2}: {role}
    Reasoning: {explanation}
  - CONSENSUS: {majority\_role}
... (for each previous round)
```

Prompt for the Self-adjustment Phase

Main Self-Adjustment Prompt Template

```
   You are {agent_name}. Based on the debate about {player_name}'s role, please
       provide your complete solution for ALL players.

GAME INFORMATION:
```

```
{game.text_game}

CURRENT FOCUS: We just finished debating {player_name}'s role.{debate_analysis}{
    previous_context}{latest_solutions_context}

You can see the conversation history above, which includes:
- Your own previous responses and solutions
- Other agents' positions and reasoning
- The debate arguments and agreements/disagreements
- Each agent's latest complete solution for all players

Based on the debate, please provide your self-adjustment solution on all players.

CRITICAL REQUIREMENTS:
 - In this self-adjustment phase, you must provide a complete assignment of roles
     for all players, not just {player_name}.
 - Do not use "unknown" or any placeholder values - you must make a definitive
     choice for each player from "knight", "knave", or "spy".

Return your complete solution in JSON format:
{
   "players": [
      {"name": "player_name", "role": "role"},
      {"name": "another_player", "role": "role"},
      ...
   ],
   "explanation": "Your reasoning after considering the debate"
}

REMINDER: Include ALL players in the "players" array, each with their assigned role.
```

Current Debate Analysis (debate_analysis)

```
   CURRENT DEBATE ANALYSIS:
- {agent_name1} (YOU): {role}
  Confidence: {confidence} (if enabled)
  Agrees with: {agent_names}
  Agree reasoning: {reasoning}
  Disagrees with: {agent_names}
  Disagree reasoning: {reasoning}
- {agent_name2}: {role}
  Confidence: {confidence} (if enabled)
  Agrees with: {agent_names}
  Agree reasoning: {reasoning}
  Disagrees with: {agent_names}
  Disagree reasoning: {reasoning}
... (for each agent in the debate)
```

Previous Debate Rounds (previous_context)

```
   PREVIOUS DEBATE ROUNDS:
Round {round_number} ({previous_player_name}):
  Debate phase:
    - {agent_name1} (YOU): {role}
      Agrees with: {agent_names}
      Agree reasoning: {reasoning}
      Disagrees with: {agent_names}
      Disagree reasoning: {reasoning}
  Self-adjustment phase:
    - {agent_name1} (YOU): {role}
      Reasoning: {explanation}
  - CONSENSUS: {majority\_role}
... (for each previous round)
```

Latest Complete Solutions (latest_solutions_context)

```
   LATEST COMPLETE SOLUTIONS FROM EACH AGENT:
- {agent_name1} (YOU):
   {player1}: {role}
   {player2}: {role}
```

```
   {player3}: {role}
   Confidence: {confidence} (if enabled)
   Reasoning: {brief_explanation}
- {agent_name2}:
   {player1}: {role}
   {player2}: {role}
   {player3}: {role}
   Confidence: {confidence} (if enabled)
   Reasoning: {brief_explanation}
... (for each agent)
```

B.3 Prompt for Final Decision

Final Discussion Prompt (for Agents)

```
   This is the FINAL DISCUSSION phase. You have access to the complete debate
       history.

GAME INFORMATION:
{game.text_game}{initial_summary}{debate_summary}

You can see the conversation history above, which includes:
- Your own responses throughout all phases (initial, debate, self-adjustment)
- Other agents' positions and reasoning from all phases
- The complete debate history and evolution of arguments

Now make your final decision for ALL players. You can reference the entire
    conversation history including your own responses and those of other agents.

CRITICAL REQUIREMENTS:
- You MUST assign each player one of the three roles: "knight", "knave", or "spy"
- Do NOT use "unknown" or any other value - you must make a definitive choice for
    each player

Return your final solution in JSON format:
{
   "players": [
      {"name": "player_name", "role": "role"},
      ...
   ],
   "explanation": "Your final comprehensive reasoning"
}
```

Initial Proposals Summary (initial_summary)

```
   INITIAL PROPOSALS:
- {agent_name1}: {player_role_assignments}
  Reasoning: {explanation}
  Confidence: {confidence} (if enabled)

- {agent_name2}: {player_role_assignments}
  Reasoning: {explanation}
  Confidence: {confidence} (if enabled)

... (for each agent)
```

Debate Rounds Summary (debate_summary)

```
   DEBATE ROUNDS AND SELF-ADJUSTMENT SUMMARY:
Round {round_number} ({player_name}):
  - {agent_name1}: {role}
    Confidence: {confidence} (if enabled)
    Agrees with: {agent_names}
    Agree reasoning: {reasoning}
    Disagrees with: {agent_names}
    Disagree reasoning: {reasoning}
  - {agent_name1} (self-adjustment): {player_role_assignments}
    Reasoning: {explanation}
    Confidence: {confidence} (if enabled)
  - CONSENSUS: {majority\_role}
```

```
... (for each round)
```

Supervisor Prompt (for Final Decision if Majority Vote not Reached)

```
   You are a supervisor AI tasked with making the final decision in a multi-agent
       debate about a Knight-Knaves-Spy game.

GAME INFORMATION:
{game.text_game}

COMPLETE DEBATE HISTORY:

INITIAL PROPOSALS:
{proposal.agent_name} initially proposed: {proposal.player_role_assignments}
Their reasoning: {proposal.explanation}

DEBATE ROUNDS:
--- Round {round_number}: {player_name} ---
{response.agent_name} thought {player_name} is a {role}
  Agreed with: {agent_names} - Reasoning: {reasoning}
  Disagreed with: {agent_names} - Reasoning: {reasoning}
Reasoning: {explanation}

SUPERVISOR INSTRUCTIONS:
As the supervisor, you have access to the complete debate history. The agents have
    been unable to reach consensus, so you must make the final decision. Consider:

1. All initial proposals and their reasoning
2. The evolution of arguments through the debate rounds
3. The consistency and logic of each agent's reasoning
4. The overall coherence of the solution
5. Any patterns or insights that emerged during the debate

Make your decision based on the most logical and well-reasoned arguments you
    observed.

Return your response in the same JSON format:
{
   "players": [
      {"name": "player_name", "role": "role"},
      ...
   ],
   "explanation": "Your final decision with comprehensive reasoning based on the
       complete debate history"
}

IMPORTANT: Keep your explanation having details but less than 100 words.
```

C Prompt Used for DeepSeek Assessment in Figure 5

```
   You are evaluating the logic soundness for an agent debate. There are {num_agents
       } agents in the debate. The debate is about a Knight-Knave-Spy game instance.
        The game setup is the following:
{{game_text}}

The debate focused on the role of the player {{player_name}}. There are three stages
    :

Stage 1: Initial proposal. Each agent propose a role deduction of this player,
    together with his reasoning.
Stage 2: Each agent decide to agree or disagree with other agents, and provide
    reasoning why he make such decisions.

Your task: to evaluate whether the agree and disagree reasons of the agent {{
    agent_name}} are sound or not. This helps us to evaluate how well the debate
    goes. To assist you, the correct solution of the game instance is:
{{ground_truth}}
```

```
In the user's input, the initial proposals of all agents and the agree and disagree
    info of the {{agent_name}} agent you are currently targeting will be provided.
    Your task is to evaluate whether these reasonings are solid enough to support
    the decision. Please follow the syllabus:
Rate 0: no applicable (i.e., no agree or disgree agents)
**For disagree reason**
- Rate 1: chaos. Show chaos in reasoning and behavior. For example, disagree with
    other agent having the same role deduction with self, but do not provide any
    logically sound argument.
- Rate 2: dangerous. Simply state the mismatch in role deduction as reason, or
    arguing that a correct reasoning from other agents is wrong.
- Rate 3: promising. Correctly point out the error from other agents as the disagree
     reason, but provide a wrong or gap argument why the other agents went wrong.
- Rate 4: solid. Correctly point out the error from other agents, and the argument
    why they are wrong is also solid.
**For agree reason**
- Rate 1: chaos. Show chaos in reasoning and behavior. For example, agree with other
     agent having the different role deduction with self, but do not provide any
    logically sound argument.
- Rate 2: dangerous. Simply state the match in role deduction as reason, without any
     further explanations. Suggesting the agent is simple compare the result rather
    than check the logic.
- Rate 3: promising. Correctly realize that the errors or limitations of self or
    other agents reasoning and behave accordingly (For example, realize self
    argument is wrong and hence agree with others), but either do not provide enough
     reasoning supports for this observation or still making wrong decisions (for
    example, agree with another wrong answer)
- Rate 4: solid. Correctly realize the errors or limitations, provide solid
    arguments, and behave correctedly (for example, realize error in reasoning but
    determine the final result is still correct, so decide to agree).

Please return your evaluation purely in the following json format:
{
   "agree_reasoning_soundness": int,
   "disagree_reasoning_soundness": int
}

Please note:
- Focus on the logic soundness for your evaluation.
- Please strictly follow the response format.
```