Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning

Anonymous ACL submission

Abstract

Multi-agent debate (MAD) has recently emerged as a promising framework for improving the reasoning performance of large language models (LLMs). Yet, whether LLM agents can genuinely engage in deliberative reasoning—beyond simple ensembling or majority voting—remains unclear. We address this question through a controlled study using the *Knight–Knave–Spy* logic puzzle, which enables precise, step-wise evaluation of debate outcomes and processes under verifiable ground truth. We systematically setup six structural and cognitive factors, including agent team size, composition, confidence visibility, debate order, debate depth, and task difficulty, to disentangle their respective effects on collective reasoning. Our results show that intrinsic reasoning strength and group diversity are the dominant drivers of debate success, while structural parameters such as order or confidence visibility offer limited gains. Beyond outcomes, process-level analyses identify key behavioral patterns: majority pressure suppresses independent correction, effective teams overturn incorrect consensus, and rational, validity-aligned reasoning most strongly predicts improvement. These findings provide valuable insights into *how* and *why* LLM debates succeed or fail, offering guidance for designing interpretable and truth-seeking multi-agent reasoning systems. Our dataset and code are available at this link.

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 really debate to advance their understanding?** To address this question meaningfully, we argue that a controlled study must examine both the debate outcome (i.e., whether accuracy improves) and the process (i.e., whether agents engage in rational interaction). Therefore, we ask two questions.

- **RQ1 (Outcome). What factors 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 significantly influence debate dynamics and solution outcomes?
- **RQ2 (Process). How do agents engage in effective debate processes?** To what extent do LLM agents actually engage in meaningful debate, such as identifying mistakes, adopting peer suggestions, or revising their answers?

To answer these questions, we adopt the *Knight–Knave–Spy* logic puzzles as a controlled reasoning environment that enables systematic measurement of both debate outcomes and debate processes. We design a multi-agent debate framework that simulates realistic deliberation: each agent makes initial proposals, debates others' reasoning, and revises its beliefs step-by-step. By systematically varying numerous factors, we quantify how each influences collective reasoning accuracy. Beyond outcomes, we further propose three desiderata for effective debate and empirically analyze how these properties correlate with performance gains. Together, these experiments provide a controlled lens on how debate structure and reasoning behavior jointly determine success or failure in multi-agent debate systems on logical reasoning. Our main contributions are as follows:

- We present a benchmark based on the *Knight–Knave–Spy* puzzle that enables rigorous evaluation of multi-agent debate on logical reasoning under verifiable ground truth.
- Through controlled studies, we systematically analyze how design choices affect debate outcomes, revealing that intrinsic model strength is the dominant factor governing debate success.
- We move beyond outcome accuracy by introducing desiderata of effective debate and analyzing the debate process, showing that process-level behaviors aligned with these properties correspond to higher correction rates and collective reasoning performance.

2 Related Work

Multi-agent debate (MAD) mitigates single-LLM limitations in multi-step reasoning, such as 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). Through adversarial and cooperative communication, MAD improves performance in tasks like mathematical reasoning (Zhang and Xiong, 2025), fact checking (Kim et al., 2024; He et al., 2025), healthcare (Lu et al., 2024; Kim et al., 2025), and code summarization (Chun et al., 2025). 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 improve debate outcomes, research focuses on structured coordination protocols. Strategies include judge-based systems for argument selection (Liang et al., 2024; Khan et al., 2024) and consensus-driven approaches like majority voting (Kaesberg et al., 2025; Li et al., 2024). Debate is further enriched by agent heterogeneity—using diverse prompts, models, or roles to reduce blind spots (Smit et al., 2024; Ye et al., 2025; Yang et al., 2025; Xing, 2025)—and confidence reporting (Lin and Hooi, 2025; Bai, 2024; Eo et al., 2025). Additionally, debate depth and turn-taking strategies significantly influence results (Lu et al., 2024; Chan et al., 2023).

Despite its promise, MAD faces fundamental challenges regarding its efficacy (Choi et al., 2025; Oh et al., 2025; Zhang et al., 2025). It remains unclear whether gains arise from interactive argumentation or simple aggregation of independent outputs (Oh et al., 2025; Estornell and Liu, 2024). Persuasiveness can eclipse accuracy, while LLM judges may introduce biases like verbosity and sycophancy (Khan et al., 2024). Homogeneous groups risk echo chambers (Bandaru et al., 2025) or premature consensus on incorrect solutions (Kaesberg et al., 2025). Moreover, increasing debate depth can entrench errors rather than improve outcomes (Ku et al., 2025), casting doubt on debate's truth-seeking capability.

A critical gap remains: while the benefits and failure modes of debate are known, the underlying interaction dynamics are largely unmeasured. This makes it difficult to distinguish process-driven reasoning from outcome aggregation. We address this through controlled experiments to unpack the factors driving outcomes and evaluate the debate process itself, elucidating *how* and *why* MAD succeeds or fails.

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 tells 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 MAD testbed for three main reasons: (i) **Reasoning challenge.** Solving these puzzles requires nontrivial logic and consistency checking, challenging even hu-

mans. This ensures agents must engage in genuine reasoning rather than rely on shallow heuristics. (ii) **Stepwise structure.** The debate progresses player-by-player, evaluating one statement at a time. This granularity allows us to isolate how intermediate decisions are influenced, unlike tasks where reasoning steps are implicit (e.g., math word problems). (iii) **Clear evaluation.** Unambiguous ground-truth solutions enable precise measurement of accuracy and error cascades. This contrasts with subjective tasks like summarization, which often have multiple valid solutions and noisier evaluations.

We constructed a dataset of 1,800 puzzles ranging from 4 to 9 players (300 per size) to probe scaling effects on difficulty and cascade risk. A four-player example is illustrated below. Further details on *rationale* and *dataset construction* are provided in Appendix B.

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 F.

- **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 framework iterates through a structured loop over all players. For each player i, two sequential subphases occur: (i) *Debate phase:* Agents discuss player i's role, referencing statements and logical consistency while engaging with peer arguments. (ii) *Self-adjustment phase:* Agents review the discussion on player i and decide whether to revise their label. This ensures in-depth analysis and localizes revisions 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 Factors Shaping Debate Outcomes

Synthesizing prior studies (Bai, 2024; Lu et al., 2024; Chan et al., 2023; Smit et al., 2024; Hegazy, 2024; Liu et al., 2024), we identify six controllable factors (Table 1) ranging from agent composition to interaction structure. Each factor drives a dedicated controlled experiment (C1–C6), detailed in Section 5.2, allowing us to systematically quantify how specific design choices determine collective reasoning success (RQ1).

4.3 Desiderata of Effective Debate Processes

Beyond measuring final accuracy, it is equally important to examine the process through which debates unfold. Outcome-based evaluations reveal *how* performance changes, but they do not explain *why* certain debates lead to genuine understanding while others fail. To uncover these mechanisms, we analyze the debate process itself and propose three desiderata that characterize the essential qualities of effective reasoning (RQ2).

Inclusive deliberation. Effective debate requires participants to respond to each other's arguments rather than speaking past one another, showing engagement by acknowledging and critically examining peers' reasoning. Moreover, minority views should not be automatically silenced by majority pressure; inclusivity ensures that valid reasoning is

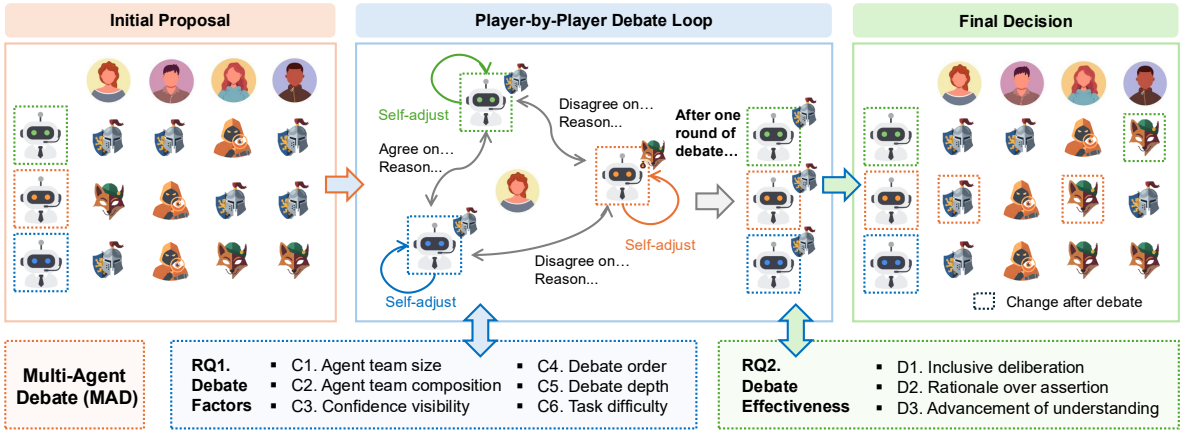

Figure 1: The illustration of the overall MAD framework design.

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.

considered even when it contradicts the dominant position. Observable signals include agents revising positions upon valid counterarguments and the adoption of minority-held correct answers.

Rationale over assertion. Debate must prioritize reasoning grounded in evidence over unsupported claims. Participants should justify their positions with verifiable information and logical consistency, not merely assert conclusions. This manifests when position changes correlate with argument validity rather than rhetorical confidence or consensus frequency.

Advancement of understanding. Finally, debate should leave participants with improved clarity or insight. Interactions that merely reinforce pre-existing biases or entrenched positions without generating new understanding are unsuccessful. Concretely, agents should correct initial errors through reasoning exchange, not merely aggregate independent guesses.

We note that these desiderata 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. Detailed empirical analysis follows in Section 6.2.

5 Experiment Design

In this section, we first describe how agent teams are constructed and define the default anchor configuration. We then outline the controlled experiments, as well as metrics and measures.

5.1 Agent Team Construction

We construct multiple agent teams to explore how different agent profiles affect debate perfor-

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 shown in Figure 2. Each dimension has three levels: ▲ High, ● Medium, and ▼ Low.

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

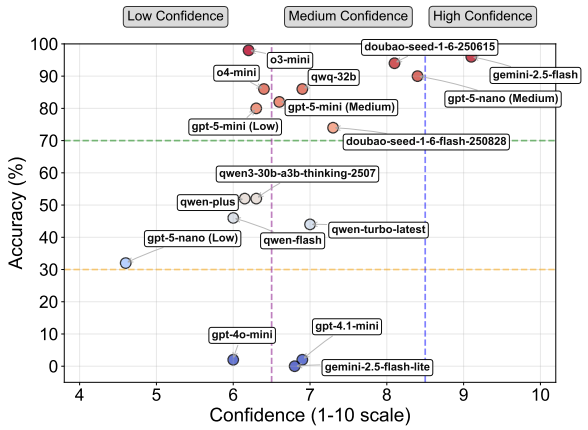

Figure 2: Single-agent accuracy versus self-reported confidence over 100 medium-difficulty games (size=6). This benchmark guides the agent team formation by categorizing models into high / medium / low performance and confidence.

mance. Each individual agent 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.

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.

5.3 Metrics and Measures

We evaluate the MAD framework using two complementary groups of measurements: (1) *outcome-level metrics* for **RQ1**, which quantify debating performance, and (2) *process-level measurements* for **RQ2**, which assess the underlying reasoning dynamics.

Outcome-level metrics. We assess *how* correctness and consensus evolve using outcome-based metrics. Specifically, we measure *strict accuracy* (all roles inferred accurately) and *smooth accuracy* (proportion of correctly identified roles per instance) to evaluate correctness at different granularities. To capture temporal dynamics, we compute the *area-under-curve (AUC)* variants for these accuracies, summarizing how rapidly agents approach the correct solution. Beyond accuracy, we evaluate alignment using *AUC agreement* (under both unanimous and majority criteria) to measure the stability of consensus formation. Detailed definitions are provided in Appendix C.

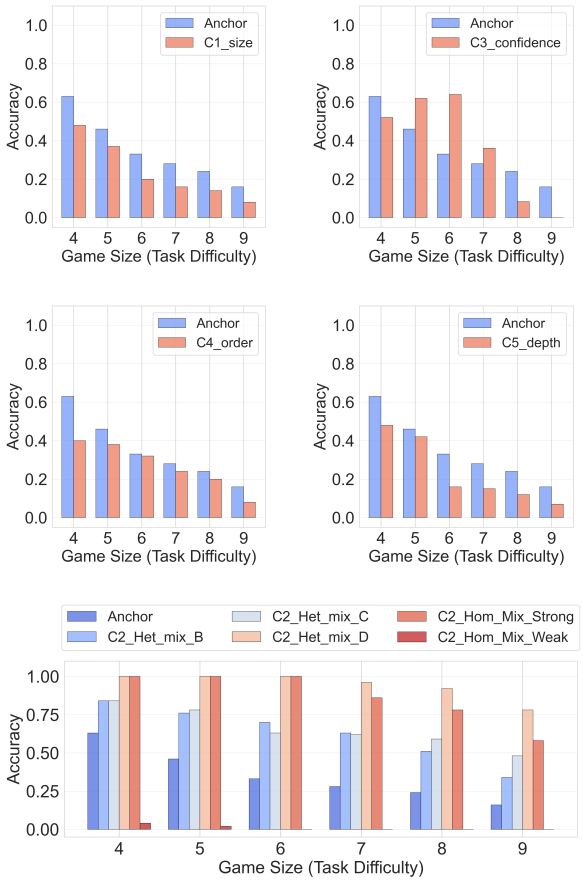

Figure 3: Overall accuracy across controlled debate settings relative to the default anchor (A). Each panel isolates one factor while keeping others fixed. The Task difficulty factor, reflected by varying game sizes, is naturally incorporated in all subfigures.

Process-level measurements. To uncover *why* debates succeed or fail, we examine internal interaction patterns aligned with the desiderata in Section 4.3. We track belief transitions across rounds, such as overturning incorrect majorities or adopting minority-held correct answers, to measure inclusive deliberation. Furthermore, we estimate *rationality-based correction rates* using an external judge model to assess whether agents revise positions in response to valid arguments rather than social conformity. These measurements operationalize our desiderata to distinguish genuine understanding from stagnation.

6 Results and Analysis

We now discuss the experiment results corresponding to our two research questions.

6.1 RQ1. What Factors Influence Debate Outcomes?

Debate becomes harder as task difficulty increases. Figure 3 summarizes the outcomes of all controlled settings relative to the default anchor configuration. Across all controlled settings in Figure 3, performance declines consistently as puzzle size grows. Larger games require longer inference chains and amplify error propagation, leading to reduced collective accuracy and agreement. This pattern holds across all debate configurations, indicating that increasing task difficulty uniformly challenges both strong and weak teams.

Table 3: Comparison of initial and post-debate accuracy across game sizes. Debate consistently improves both instance- and agent-level performance, with larger gains in simpler tasks.

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**

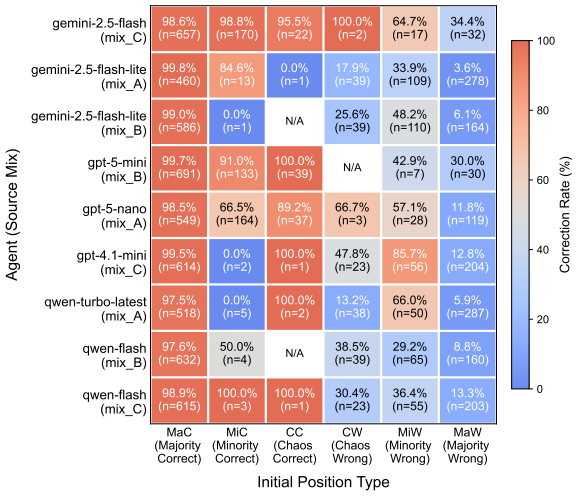

Figure 4: Heatmap on initial states. x-axis: initial stats. Here Ma = majority, Mi = Minorty, first C = chaos, second C = correct, and W = wrong. So MaC means the model starts with a majority and correct initial position in a debate round. y-axis: different models. This figure counts for each agent their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

Debate success is primarily driven by reasoning strength and diversity. As shown in Figure 3 (bottom), base model strength governs performance: *Hom-Mix D (strong)* consistently outperforms *Hom-Mix Weak*, establishing a strength-dependent ceiling. Among heterogeneous teams, *Het-Mix D (strong)* leads, followed by *Het-Mix A (balance)*, while weaker mixes like *Het-Mix B (stress)* lag behind. Comparisons between *Hom-Mix Strong* and *Het-Mix D* indicate that diversity offers modest, consistent gains when strong rea-

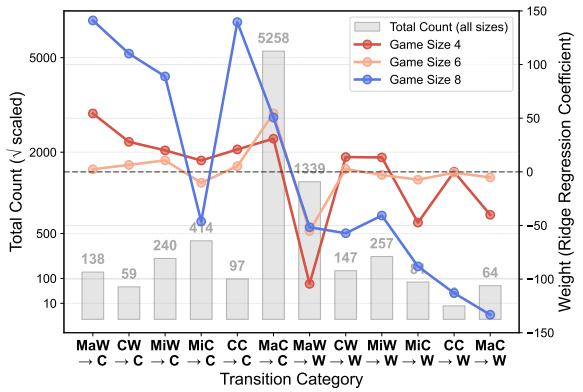

Figure 5: Correlations between states transition and final accuracy of the game instance. x-axis: 12 state transition possibilities. Left y-axis: counts of state transitions. Right y-axis, the weight each state transition contributes to the final accuracy of a game instance: for each game instance we 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.

soners are present. In contrast, purely weak teams (*Hom-Mix Weak*) see negligible benefit from structural changes like order or depth. 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.

Debate consistently raises collective accuracy within each task. As part of our outcome-level evaluation (RQ1), Table 3 compares performance before and after debate across different game sizes. Instance-level accuracy—measured by majority voting per player and verified against ground truth—shows substantial improvement, increasing by 52% (size-4) and 32% (size-8). A similar upward trend is observed at the agent level, indicating that individual participants also benefit from collective deliberation. These findings confirm that the debate mechanism itself leads to measurable performance gains within each task, even as overall accuracy remains bounded by task complexity and individual reasoning strength.

Key predictors of debate success. To further examine which factors most strongly predict outcome metrics, we perform regression analyses (Tables 4 and 5 in Appendix D) and identify three primary drivers We further identify three key observations: (1) **initial accuracy and team size are the strongest predictors of debate success**; (2) **moderate initial variability fosters productive adaptation**; and (3) **team composition and initial balance shape the quality of consensus**. Due to space constraints, we defer the detailed analyses to Appendix D.

6.2 RQ2. How Do Agents Engage in Effective Debate Processes?

Majority pressure suppresses agents' independent correction. We first analyze agent behavior based on their initial position (Figure 4) to test the "Inclusive deliberation" dimension for debate process. Agents starting in a correct majority (MaC) are exceptionally stable (>97% retention). However, facing an incorrect majority (MaW) reveals a clear capability gap: stronger models like *gemini-2.5-flash* and *gpt-5-mini* show moderate correction abilities (34.4% and 30.0% respectively), whereas weaker models are almost entirely swayed, with *gemini-2.5-flash-lite* (mix_A) correcting only 3.6% of cases. This indicates weaker agents defer to consensus over evidence.

Further investigation reveals a "minority correction asymmetry": agents more readily maintain a correct minority position (MiC) than correct an incorrect one (MiW), indicating a bias toward initial stances. We also observe teammate effects, where agents like *gemini-2.5-flash-lite* show different correction rates across mixes. This confirms that debate is path-dependent and contingent on team composition, not just individual capability.

Debate success depends on agents' ability to overturn incorrect consensus. We analyze belief-state transitions and their impact on overall accuracy (Figure 5). Using ridge regression, we estimate the effect of each transition type on instance-level smooth accuracy. The distribution of transition counts (gray bars) indicates that debates are dominated by stable or consensus-preserving states (e.g., MaC→C, CC→C), while majority-reversal transitions (MaW→C) are relatively rare. Despite their low frequency, these reversals have the strongest positive coefficients, showing that correcting an incorrect consensus contributes most to final accuracy.

As difficulty increases (red, orange, and blue curves for game sizes 4, 6, and 8), the positive effect of MaW→C transitions grows sharply, while transitions like MaC→W become increasingly detrimental. This pattern suggests that effective debates rely less on preserving majority agree-

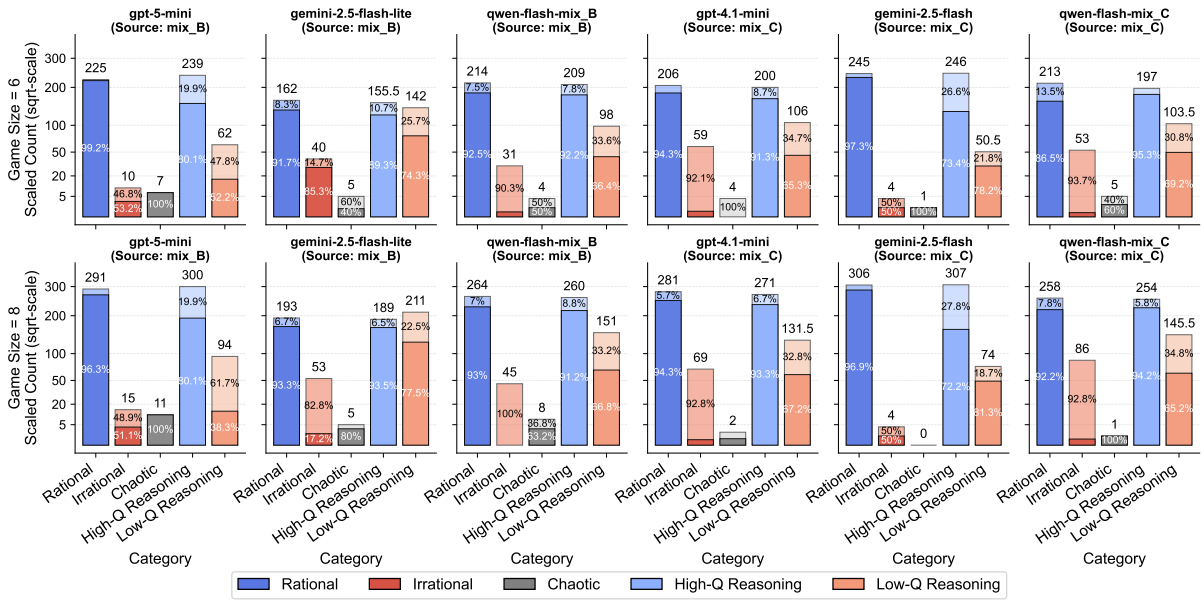

Figure 6: Whether agents behaves rationally. The reasonings in the debate round for each agent is sent to deepseek-r1-0528 model to evaluate (1 through 4). Then we count (in the first and second bar) if agent behaves rationally or not: whether agent follows high (3 or 4) rating opinions from either himself 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.

ment and more on revising incorrect consensus through minority-informed reasoning. Progress arises not from conformity, but from reflective self-correction in response to counterarguments.

Rational, validity-aligned reasoning drives accurate correction. We assess debate process by evaluating agents by whether they follows the rationales themselves or provide quality opinions for others given 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) (Figure 6). **A human evaluation in Appendix E supports this validation.** For individual agents, rationality—following high-quality opinions or aligning with valid arguments—predicts success. Rational behaviors yield correction rates exceeding 90%, whereas irrational ones drop below 55% (despite *gemini-2.5-flash-lite*), demonstrating that receptiveness to valid counterarguments is critical for error correction.

As opinion providers, a "strong model paradox" emerges: teammates correct less often when strong models like *gpt-5-mini (Medium)* provide high-quality reasoning (80% in size 8) compared to weaker providers like *gpt-4.1-mini* (93%). This reveals that while strong agents provide high-quality opinions, weaker teammates lack the capacity to assess them (Figure 3), leading to irrational behavior that undermines the debate.

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 limited to the LLMs available at the time of study. As model capabilities evolve, future systems with stronger reasoning or calibration may exhibit different debate dynamics. We also test a narrow range of team compositions and debate protocols; alternative designs, such as asynchronous deliberation, dynamic roles, or adversarial setups, remain open for exploration.

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.

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 Rationale for choosing Knight-Knave-Spy and Dataset Construction

Why KKS is an appropriate testbed for analyzing debate dynamics? We have described several reasons in Section 3. Our focus is on logic-based reasoning under unambiguous ground truth, and KKS uniquely satisfies this requirement:

- **Unambiguous and definitive solutions**: Every game instance is guaranteed to have a single correct assignment of roles, providing a clean and objective ground truth. This avoids the ambiguity found in tasks like open-ended QA or summarization.
- **Granular, multi-step evaluation (more than a final answer)**: Unlike math datasets that reduce evaluation to one number, KKS allows us to evaluate role deductions for every player, enabling fine-grained analysis of debate trajectories, per-player adoption, and intermediate error propagation.
- **Decomposable logic aligned with debate structure**: Many statements produce local logical consequences (e.g., "I am a knave" $\rightarrow$ must be spy), which then interact compositionally across the puzzle. This naturally aligns with the player-by-player debate loop used in MAD and allows us to examine how agents build or revise partial reasoning over multiple rounds.
- **Flexible but rigorously controlled complexity**: By adjusting game size, we can scale difficulty systematically. The structure is flexible yet fully rigid, letting us generate many instances that are comparable and logically sound.

These properties provide the controlled, interpretable conditions needed to isolate the effects of debate design, which satisfy the goal of our paper.

Dataset construction and validation As noted in Section 3, we construct 1,800 puzzles (300 each for game sizes 4–9). All instances are generated and validated entirely via symbolic code—no LLM is involved—which guarantees logical consistency. We detail the **game rule** and **game instance generation** below.

Game rule.

- **Three roles:**
  - **Knight:** always telling the truth.
  - **Knave:** always lying.
  - **Spy:** may either tell the truth or lie.
- **Game setup:**
  - In the current dataset, every instance contains *exactly one* spy, and this is known to the model.
  - Each player is assigned a role and provides a statement, whose truthfulness must align with the player's role.
  - Each game instance is guaranteed to have a unique correct solution.
  - The model's task is to identify the unique solution by deducing the roles of all players.

Game instance generation.

- **Symbolic generation and validation.**
  - Players and roles are represented abstractly by indices.
  - We design a small set of statement templates, including: (i) membership of an index in a list, (ii) parity of the length of a list, and (iii) logical compositions using `and`, `or`, and `not`.
  - To prevent cyclic dependencies, the statement of player i may only reference players with indices $< i$.

- We then randomly generate roles and statements by: random role assignment, random template selection, and random construction of the involved player list under the chosen template.
- Finally, we apply a brute-force enumeration algorithm to ensure there exists a *unique* role assignment consistent with all statements and their role-conditioned truthfulness. For a size-8 instance, this requires enumerating 3^8 candidate assignments, which is computationally affordable.

- **Text augmentation.**
  - After symbolic generation and validation, we render a human-readable instance via a script.
  - **Name rendering:** we sample names from a fixed pool and randomly assign them to players.
  - **Template rendering:** for each symbolic template, we pre-define its grammar and deterministically translate the symbolic statement into natural language.

- **Important note.** We use only code throughout the entire instance generation process, with no LLM-based component involved, thereby guaranteeing the logical rigidity of the KKS dataset.

C Outcome-level Quantitative Metrics

Final Output Accuracy. By default we report **strict accuracy** over instances:

$$\text{StrictAccuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\hat{y}_i = y_i], \quad (1)$$

where $\hat{y}_i$ is the final predicted assignment for instance i and y_i is ground truth. We optionally report a **smooth** variant that averages per-player correctness within each instance:

$$\text{SmoothAccuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_i}\sum_{p=1}^{P_i}\mathbf{1}[\hat{y}_{i,p} = y_{i,p}]. \quad (2)$$

AUC-Strict-Accuracies. To summarize how strict correctness evolves during debate, we define the area-under-curve of strict accuracy across rounds:

$$\text{AUC\_Strict} = \frac{1}{T}\sum_{t=1}^{T}\text{StrictAccuracy}(t), \quad (3)$$

where $\text{StrictAccuracy}(t)$ equals 1 if, at round t, the round-wise majority prediction matches ground truth for *all* players in an instance, and 0 otherwise. Values are in $[0, 1]$; higher is earlier and more sustained correctness.

AUC-Smooth-Accuracies. Analogously, we aggregate the per-round smooth accuracy:

$$\text{AUC\_Smooth} = \frac{1}{T}\sum_{t=1}^{T}\text{SmoothAccuracy}(t), \quad (4)$$

where $\text{SmoothAccuracy}(t)$ is the proportion of players correctly predicted by the round-t majority. Values are in $[0, 1]$.

AUC-Agreement-All. We measure how often agents unanimously agree on player roles during the debate:

$$\text{AUC\_Agree-All} = \frac{1}{T}\sum_{t=1}^{T}\text{AgreeAll}(t), \quad (5)$$

$$\text{AgreeAll}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_i}\sum_{p=1}^{P_i}\mathbf{1}[\mathsf{All}(p,t)], \quad (6)$$

where $\mathsf{All}(p, t)$ indicates that all agents agree on player p at round t. Values are in $[0, 1]$ and reflect unanimous player-level alignment over time.

AUC-Agreement-Major. We also measure majority agreement (at least half of agents agree) over time:

$$\text{AUC\_Agree-Major} = \frac{1}{T}\sum_{t=1}^{T}\text{AgreeMajor}(t), \quad (7)$$

$$\text{AgreeMajor}(t) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{P_i}\sum_{p=1}^{P_i}\mathbf{1}[\mathsf{Major}(i,p,t)], \quad (8)$$

where $\mathsf{Major}(i, p, t) := \max_\ell \#\{a : \hat{y}_{i,p,a,t} = \ell\} \geq \lceil A/2 \rceil$, denoting the event that, at round t, at least half of the agents assign the same label to player p in instance i. A is the number of agents. Values are in $[0, 1]$; higher indicates earlier and more sustained majority alignment.

D Regression Analysis of Factors Predicting Outcome Metrics

In addition to the six controlled factors defined in Section 4.2, we include a derived variable, *initial chaos*, which captures variability in agents' first-round predictions before debate begins. A game instance is labeled as chaotic when no clear majority exists in the initial role assignments, reflecting early disagreement that may influence subsequent deliberation. Table 4 (accuracy) and Table 5 (agreement) summarize the results. Overall, *initial accuracy*, *agent count*, and moderate *initial chaos* positively contribute to performance, whereas increasing *task difficulty* and unbalanced *agent team composition* reduce agreement stability.

Initial accuracy and team size are the strongest predictors of debate success. The regression on *final smooth accuracy* (Table 4) explains about 39% of the variance ($R^2 = 0.393$), indicating a solid model fit. The most influential predictor is *initial smooth accuracy* ($p < 0.001$), with a large positive coefficient ($\beta = 0.600$): teams starting from stronger baselines achieve higher post-debate accuracy. The *number of agents* also has a significant positive effect ($p < 0.001$), suggesting that larger collectives promote improved aggregate reasoning. In contrast, *game size* shows a negative association ($p < 0.001$), confirming that increasing task complexity hinders accuracy. Other individual-level variables and majority/minority indicators are not significant, implying that performance gains arise mainly from baseline quality and group scale rather than agent positioning.

Moderate initial variability stimulates productive adaptation. The presence of *initial chaos* ($p < 0.001$) increases final accuracy, indicating that some early disagreement can promote effective deliberation. Debate depth, confidence visibility, and debate order remain insignificant, underscoring the robustness of the core structural factors. Although detailed mixture coefficients are omitted, several heterogeneous settings (e.g., Mix B and Mix C) show positive effects, highlighting that team composition also contributes to accuracy improvements.

Team composition and initial balance shape consensus quality. The regression on *AUC agreement* (Table 5) explains approximately 29% of the variance ($R^2 = 0.290$), reflecting a moderate fit. *Game size* ($p < 0.001$) and *number of agents* ($p < 0.001$) positively predict agreement, showing that larger, more populated debates reach steadier consensus. Similarly, *initial smooth accuracy* ($p < 0.01$) predicts higher agreement levels. However, when strong or weak agents begin in minority positions, agreement quality declines significantly ($p < 0.001$), indicating that balanced influence within teams supports stable convergence. Conversely, excessive *initial chaos* ($p < 0.001$) lowers agreement, suggesting that while moderate variability aids adaptation, too much early disorder hinders group alignment. As before, balanced heterogeneous mixes (e.g., Mix C, Mix H) yield higher agreement scores, reinforcing that diversity, when structured, is beneficial.

Table 4: Regression analysis of factors predicting final output accuracy. Higher initial accuracy, larger teams, and moderate initial chaos enhance performance, while greater task difficulty reduces it.

Variable	Coef.	Std. Err.	Sig.
Constant	0.044	0.013	**
Game size	−0.030	0.007	***
# of agents	0.066	0.014	***
Debate depth	0.019	0.015	
Init. smooth acc.	0.600	0.100	***
Init. strict acc.	−0.015	0.048	
Strong init. smooth acc.	0.170	0.091	
Strong init. strict acc.	0.005	0.057	
Weak init. smooth acc.	0.018	0.087	
Weak init. strict acc.	0.020	0.069	
Strong in minority	0.011	0.012	
Weak in minority	0.003	0.014	
Conf. visible (T)	0.023	0.017	
Debate order same (T)	0.035	0.019	
Init. has chaos (T)	0.085	0.020	***
R^2	0.393		
Adj. R^2	0.380		
Observations	745		

Note. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Coefficients represent OLS estimates. Variables marked "(T)" are binary indicators (True = 1).

E Human Judging as a Complement to LLM-as-a-Judge

To validate the rationality of the *LLM-as-a-judge*, we conduct a manual annotation study on a randomly selected subset of 100 debate instances from our dataset. Two independent human annotators (authors of this paper) are tasked with rating the arguments using a simplified version of the same 1–4 scale employed by the DEEPSEEK-R1 judge. For each instance, we manually evaluate the soundness of both the *agree* and *disagree* rea-

Table 5: Regression analysis of factors predicting agreement stability. Balanced team composition and moderate variability foster consensus, whereas excessive initial chaos or imbalance hinder it.

Variable	Coef.	Std. Err.	Sig.
Constant	0.040	0.012	**
Game size	0.052	0.006	***
# of agents	0.111	0.012	***
Debate depth	0.019	0.013	
Init. smooth acc.	0.275	0.087	**
Init. strict acc.	0.018	0.042	
Strong init. smooth acc.	−0.022	0.079	
Strong init. strict acc.	−0.025	0.049	
Weak init. smooth acc.	−0.032	0.076	
Weak init. strict acc.	0.067	0.060	
Strong in minority	−0.102	0.010	***
Weak in minority	−0.055	0.013	***
Confidence visible (T)	0.025	0.015	
Debate order same (T)	0.002	0.017	
Initial has chaos (T)	−0.117	0.017	***
R^2		0.290	
Adj. R^2		0.274	
Observations		745	

Note. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Coefficients represent OLS estimates. Variables marked "(T)" are binary indicators (True = 1).

soning, resulting in 200 judgments (100 instances $\times$ 2).

Rating Scale. The rating scale used for both human annotators and the judge is defined as follows:

- **Rating 1**: Reasoning exhibits logical contradictions or provides no sound argument.
- **Rating 2**: Reasoning merely states role mismatch/match, or incorrectly argues that a correct peer reasoning is wrong.
- **Rating 3**: Reasoning correctly identifies a peer's error or limitation, but the counter-argument is flawed or insufficient.
- **Rating 4**: Reasoning correctly identifies errors or limitations and provides a solid, verifiable counter-argument or justification.

We use Krippendorff's α to measure both inter-coder agreement and human–machine agreement. Krippendorff's α is a reliability coefficient defined as the ratio of observed disagreement relative to the disagreement expected by chance:

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{9}$$

Table 6: Krippendorff's α for human–machine and inter-annotator agreement.

Comparison	Krippendorff's α
Machine vs. Annotator 1	0.9527
Machine vs. Annotator 2	0.9026
Annotator 1 vs. Annotator 2	0.8829

where D_o denotes the observed disagreement and D_e denotes the expected disagreement by chance.

The agreement metrics between the DEEPSEEK-R1 judge and the two human annotators are reported in Table 6. Overall, the high agreement scores (**Krippendorff's** $\alpha > 0.8$) validate the conclusions drawn from our process-level analyses (RQ2), particularly our findings on rational versus irrational agent behavior and the value of overturning incorrect consensus (Figure 5 and Figure 6).

F Prompt Design of the Multi-Agent Debate System

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

F.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.
```

F.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)
```

F.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.
```

G Prompt Used for DeepSeek Assessment in Figure 6

```
   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.
```