SVR-MAD: A Bayesian-Inspired Framework for Posterior-Guided Multi-Agent Debate

Weifan Jiang, Rana Shahout, Minghao Li, Zhenting Qi,
Yilun Du, Michael Mitzenmacher, Minlan Yu
Harvard University
Correspondence: weifanjiang@g.harvard.edu

Abstract

Multi-Agent Debate (MAD) improves LLM-agent accuracy but suffers from rapid context growth, limiting scalability in larger multi-agent settings. Existing methods prune low-utility communications using prior signals, such as token-level log-likelihoods or LLM self-reported confidence. However, these signals become unreliable under hallucination, degrading the accuracy of MAD methods that rely on them. We propose SVR-MAD, a Bayesian-inspired MAD framework that treats pre-debate signals as priors and debate outcomes as posterior-style evidence for estimating agent correctness. SVR-MAD uses this evidence to incrementally construct the communication graph, prioritizing agents whose answers survive peer challenges. Experiments across multiple LLMs and benchmarks show that SVR-MAD reduces token cost by up to 61% while matching or improving accuracy relative to the most accurate competing MAD baseline.

1 Introduction

Multi-Agent Debate (MAD) improves the accuracy of Large Language Model (LLM) agents across diverse domains, including research, mathematics, and coding (Du et al., 2024; Li et al., 2023; Wu et al., 2024). By exchanging intermediate reasoning steps and answers across rounds, agents can revise their responses based on peer information and achieve accuracy gains beyond single-agent settings. However, MAD becomes costly as the communication graph grows. Each shared message is appended to another agent's context, increasing the input-token cost and KV cache usage. With more agents or debate rounds, this overhead compounds quickly, limiting scalability.

Recent work reduces the cost of MAD by removing nodes (i.e., participating agents) and/or edges (i.e., communication links between agents) from all-to-all communication graphs (Liu et al., 2026; Chen et al., 2026; Zeng et al., 2025; Zhu et al., 2026). These methods typically leverage information available prior to the debate (e.g., token-level log-likelihoods, LLM's self-stated confidence) to identify and remove potentially low-utility communications. For example, outputs with high minimum log-likelihood values are often considered to be reliable, and therefore are likely to be correct without further debate (Chen et al., 2026).

In this work, we uncover that although pre-debate signals are effective in common cases, their informativeness can collapse on challenging problem instances relative to the LLM's capabilities. On such instances, incorrect agents may hallucinate with high confidence, while correct agents may remain uncertain (Kalai et al., 2025). Methods based on prior signals can prematurely exclude agents or communication links that would otherwise be beneficial, thereby reducing MAD's accuracy gains.

To address this limitation, we take a Bayesian-inspired view of MAD: prior agent signals can be unreliable under hallucination, whereas debate outcomes provide posterior evidence of reliability. We focus on whether an agent preserves its answer after incorporating peer messages. Intuitively, an agent supported by sound reasoning is less likely to be convinced by flawed peer arguments. We examine the *Survival Rate* (SVR), which measures how often an agent retains its original belief after being challenged. Because SVR is grounded in observed post-debate behavior, it provides a more hallucination-robust signal for identifying trustworthy reasoning processes and answers.

We propose SVR-MAD, an efficient MAD framework that incrementally builds a communication graph using posterior debate outcomes. SVR-MAD initializes agent correctness scores from prior signals, probes S communication links to high-prior receivers, and updates scores based on whether agents retain or revise their answers. It then prioritizes high-score agents in later commu-

nications and terminates once an agent's score exceeds a threshold. This greedy strategy identifies reliable solutions early, reducing token costs. Across two LLMs and datasets, SVR-MAD reduces token costs by up to 61% while matching or improving accuracy relative to the most accurate competing MAD baseline.

2 Background and Related Works

MAD preliminary. In MAD, LLM agents iteratively exchange progress. Suppose there are N participating agents. Let C_t^i be the output (i.e., reasoning and answer) from agent i at turn t. Then, in all-to-all MAD, agent i generates the next-turn output, C_{t+1}^i, based on the following context:

[Question] + [turn $1, \ldots, t-1$ history] + C_t^i + $DebateFormat(C_t^1, \ldots, C_t^{i-1}, C_t^{i+1} \ldots C_t^N)$

where $DebateFormat$ is a string manipulation function that concatenates peer outputs and adds a prompting hint to guide agent i to reflect on peer contexts, such as *"These are the solutions from other agents [Insert peer solutions here]...Using the solutions from other agents as additional information, provide your answer..."* (Du et al., 2024).

While MAD improves LLM-agent accuracy, it scales poorly with more agents due to rapid context growth. Under all-to-all communication, context size grows as $\mathcal{O}(N^2 t)$, increasing computation and KV-cache pressure while also risking long-context degradation from diluted salient information (Liu et al., 2024).

Efficient MAD solutions. Existing works leverage pre-debate signals to improve MAD. One line of work uses agent-level reliability signals to identify agents whose answers already appear reliable and who can therefore conclude without any debate (Chen et al., 2026; Eo et al., 2025). Another line uses cross-agent similarity signals to remove communication links that appear redundant (Zeng et al., 2025; Zhu et al., 2026).

3 Analysis of Prior and Posterior Signals

Prior signals. Prior signals are derived from agents' initial responses. They typically measure output quality and are used to exclude agents whose initial responses already appear reliable from debate, and/or prune the communication graph to prioritize deliberation with high-quality agents, aiming to improve accuracy with less communication.

We study three prior signals: minimum log-likelihood (min-LL), which captures the least likely token; perplexity (PPL), which measures response-level generation uncertainty; and the LLM's self-reported confidence (conf).

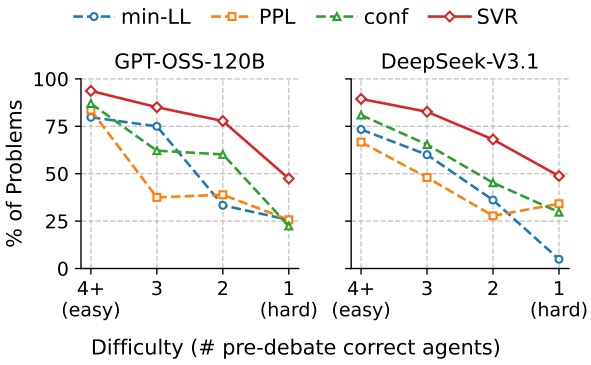

Figure 1: **SVR better identifies correct agents than prior signals under increasing difficulty.** Percentage of problems where the top-ranked agent under each signal is correct, as a function of increasing difficulty.

Our proposal: SVR. SVR measures how often an agent retains its initial belief after debating with peers. This evidence-grounded signal provides a posterior estimate of agent reliability that is less dependent on potentially hallucinated prior information. Let D, r, and c denote the total number of debates involving an agent as receiver and the number of times the agent retains or changes its belief after debate, respectively. We define $SVR = (r-c)/D$.

Approach. We validate SVR on two LLMs, GPT-OSS-120B (OpenAI, 2025) and DeepSeek-V3.1 (DeepSeek-AI et al., 2025), using IMO-AnswerBench (Luong et al., 2025). For each question, we run six agents and probe debate outcomes for all disagreeing pairs. We group questions by difficulty, defined by the number of agents correct before debate, and compare SVR with three prior signals by measuring how often the top-ranked agent under each signal is correct.

Results. Figure 1 reports this percentage across difficulty groups. From 4+ (easiest) to 1 correct agent (hardest), the correctness of top-ranked agents under min-LL, PPL, and conf drops from 67–87% to 5–34%. **This sharp decline shows that prior signals become unreliable on difficult problems, where hallucinated reasoning makes model behavior poorly aligned with correctness.** In contrast, SVR remains predictive at high difficulty: among questions with only 2 correct agents, SVR achieves 77.8% and 68.1% on OSS and V3.1, outperforming the strongest prior by 17.6 and 22.8 percentage points; with only 1 correct agent, SVR achieves 47.4% and 48.8%, outperforming the

strongest prior by 21.8 and 14.6 percentage points.

4 SVR-MAD design

Algorithm 1 SVR-MAD

Input: (1) Agents $\mathcal{A}_1 \dots \mathcal{A}_N$, (2) Pre-debate answers $y_1 \dots y_N$, (3) Communications per round S, (4) Budget B_{max}, (5) Acceptance threshold τ

Output: Final answer $\hat{y}$

1: **Initialize** correctness scores $Corr[\mathcal{A}_i] = Prior(\mathcal{A}_i)$ for $1 \le i \le N$
2: **Initialize** $B = \lfloor B_{\max}/S \rfloor \cdot S$
3: **Initialize** $Debated[\mathcal{A}_i] = \emptyset$ for $1 \le i \le N$
4: **while** $B > 0$ **do**
5: $\quad \mathcal{A}_r = \arg\max_{\mathcal{A}} Corr[\mathcal{A}]$
6: $\quad \mathcal{P} = \{\mathcal{A}_j : y_j \neq y_r, \mathcal{A}_j \notin Debated[\mathcal{A}_r]\}$
7: $\quad$ **for** $\mathcal{A}_s \in$ top-S of $\mathcal{P}$ ranked by $Prior$ or $Corr$ **do**
8: $\quad\quad Outcome = Debate(\mathcal{A}_s \to \mathcal{A}_r)$
9: $\quad\quad$ Update $Corr[\mathcal{A}_r]$ by $Outcome$
10: $\quad\quad$ Add $\mathcal{A}_s$ to $Debated[\mathcal{A}_r]$
11: $\quad\quad B \mathrel{-}= 1$
12: $\quad$ **end for**
13: $\quad$ **if** $Corr[\mathcal{A}_r] \ge \tau$ **then**
14: $\quad\quad$ **return** $\hat{y} = y_r$
15: $\quad$ **end if**
16: **end while**
17: **return** Fallback majority-voting

Algorithm 1 presents SVR-MAD's workflow. The inputs are the pre-debate states of all agents, including prior signals, reasoning tokens, and initial answers. SVR-MAD has the following steps:

Prior-based correctness score initialization (line 1): We initialize each agent's correctness score using prior signals, which provide an initial estimate of whether $\mathcal{A}_i$ holds a correct answer. These scores are updated as debate outcomes become available.

Incremental graph construction (line 4–16): SVR-MAD operates iteratively by selecting one agent (A_r) as the receiver at a time. It then pairs the receiver with S peers and performs S pairwise debates in parallel. The outcomes (i.e., whether A_r retained or revised its belief after debate) are used to update A_r's correctness score.

Sender and receiver selection (line 5–7): In each iteration, SVR-MAD selects the agent with the highest correctness score as the receiver (excluding those with all disagreeing peers already probed). This greedy strategy prioritizes agents most likely to be correct, helping SVR-MAD confirm reliable solutions and terminate with fewer communications. Among peers who disagree and haven't debated with $\mathcal{A}_r$, SVR-MAD selects up to S most challenging senders by their $Corr$ or $Prior$ scores. $Corr$ prioritizes senders with evidence-verified arguments, while $Prior$ favors ones appearing to be sound. Appendix A.3 ablates these two scores.

Posterior-style score update (line 8–11): SVR-MAD updates $\mathcal{A}_i$'s correctness score based on debate outcomes. Let D, r, and c denote the running counts of observed debates, retentions, and changes, respectively. After each debate, we increment $D \mathrel{+}= 1$ and either $r \mathrel{+}= 1$ or $c \mathrel{+}= 1$. We then apply a posterior-dominant rule:

$$Corr[A_r] \quad = \quad \begin{cases} SVR(D, r, c), & \text{if } D > 0, \\ Prior(A_r), & \text{otherwise.} \end{cases}$$

Thus, the prior only initializes $Corr$ and is overwritten by the observed SVR after a single debate.

Early termination (line 13–15): If the receiver's updated correctness score qualifies for acceptance, SVR-MAD accepts its answer and terminates the debate early to reduce overall communication cost. Note that an agent with no debate participation, i.e., $Corr = Prior$, does not qualify for acceptance.

Fallback majority voting (line 17): If no agent commits within budget, each agent casts one vote using the majority of its observed post-debate answers, with its pre-debate answer y_i as the tiebreaker. SVR-MAD returns the majority over these N votes, breaking ties by the pre-debate majority among $y_1, \dots, y_N$.

Disclaimer: SVR-MAD is not a formal Bayesian method. Existing efficient MAD solutions prune communication based on pre-debate signals, while SVR-MAD's Bayesian inspiration is reflected by Algorithm 1's score update mechanism rather than a formal Bayesian posterior.

5 Evaluation

Baselines. We compare SVR-MAD to (1) Self Consistency (Wang et al., 2023), which aggregates independent agent solutions by majority vote without debate; (2) GroupDebate (Liu et al., 2026), which restricts full-context exchange to random subgroups and allows only highly-compressed messages (e.g., final answers) across groups; (3) SID-ET (Chen et al., 2026), which uses token log-likelihood-based self-confidence to select agents for debate; and (4) S^2-MAD (Zeng et al., 2025), which prunes communications between similar-

LLM	Method	IMO-AnswerBench			HLE		
		NComm ↓	Tok ($\times 10^3$) ↓	Acc ↑	NComm ↓	Tok ($\times 10^3$) ↓	Acc ↑
GPT-OSS-120B	Self Consistency	0.0	45.4	36.7	0.0	14.1	13.9
	GroupDebate	19.4	164.1	41.8	18.4	111.0	20.2
	SID-ET	17.5	85.5	38.7	12.8	39.6	17.2
	S^2-MAD	22.8	131.4	42.1	14.0	67.7	21.0
	SVR-MAD	**5.6**	**82.1**	**42.8**	**7.3**	**39.1**	**21.4**
DeepSeek-V3.1	Self Consistency	0.0	20.2	31.8	0.0	9.8	14.1
	GroupDebate	19.8	234.6	40.1	19.6	108.7	14.9
	SID-ET	20.0	103.0	33.8	16.6	33.7	14.1
	S^2-MAD	20.9	154.7	36.8	18.7	71.7	**16.7**
	SVR-MAD	**8.9**	**91.7**	**41.5**	**5.3**	**30.3**	**16.7**

Table 1: **SVR-MAD achieves the best cost-accuracy tradeoff across all evaluated settings.** Evaluation results across two datasets and two LLMs. Arrows indicate whether higher (↑) or lower (↓) metric values are better. Best results are bold (NComm/Tok comparisons exclude Self Consistency, which is a no-debate reference).

answer agents. All methods use the same pre-debate responses per question.

Metrics. We capture the tradeoff between communication cost and accuracy using three metrics: (1) number of communications (NComm), which counts pairwise context transfers in the abstract communication graph; (2) total tokens, which captures the concrete overhead under specific model, dataset, and prompt settings; and (3) accuracy.

MAD setting. We use 6 agents per problem. All methods run for at most two debate rounds and may terminate after round one if at least five agents reach consensus, avoiding cost inflation.

LLMs and datasets. We evaluate GPT-OSS-120B and DeepSeek-V3.1 on IMO-AnswerBench and a curated subset of HLE (Phan et al., 2026). For HLE, we use text-only[1], and closed-ended questions for deterministic correctness evaluation. We use 7/8 subjects in HLE, excluding math due to overlapped domain with IMO. We exclude questions where all six agents produce the same pre-debate answer, as debate is unnecessary.

Additional experimental settings. We detail remaining experimental settings in Appendix A.

5.1 Main results

SVR-MAD achieves the best cost-accuracy tradeoff across all evaluated settings. SVR-MAD matches or exceeds every MAD baseline on accuracy while reducing per-question communications by 48–75% and total tokens by 38–61%, relative to the most accurate MAD baseline. On three of the four settings, SVR-MAD strictly improves accuracy, with gains of up to 1.8, 7.7, and 4.7 percentage points compared to GroupDebate, SID-ET, and S^2-MAD, respectively. On V3.1/HLE, SVR-MAD matches S^2-MAD's accuracy but uses 72%/58% fewer communications/tokens per question.

[1]Both evaluated LLMs require text-only inputs.

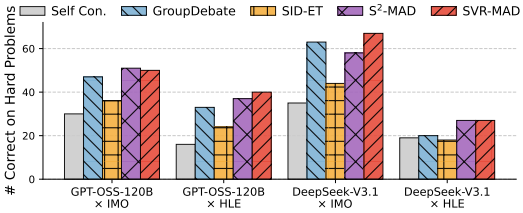

Figure 2: **SVR-MAD achieves high accuracy on hard problems across all evaluated settings.** Accuracy of each method on problems where the number of correct agents before debate is within 1-3.

SVR-MAD achieves robust accuracy on hard problems while remaining cost-efficient. Figure 2 shows the number of hard problems each method answers correctly, where hard problems have 1–3 correct agents before debate. SVR-MAD matches or exceeds all baselines on three of four settings, trailing only by one question to S^2-MAD on OSS/IMO. As S^2-MAD incurs 1.4–2.5× higher token cost on these hard problems, SVR-MAD remains the most robust and cost-efficient.

Appendix B assesses the statistical uncertainty of Table 1's results. In conclusion, SVR-MAD either significantly improves accuracy or shows no significant accuracy difference while consistently reducing communication costs.

5.2 Analysis

5.2.1 SVR Ablation

We evaluate whether SVR itself provides a better post-debate signal than alternative confidence mea-

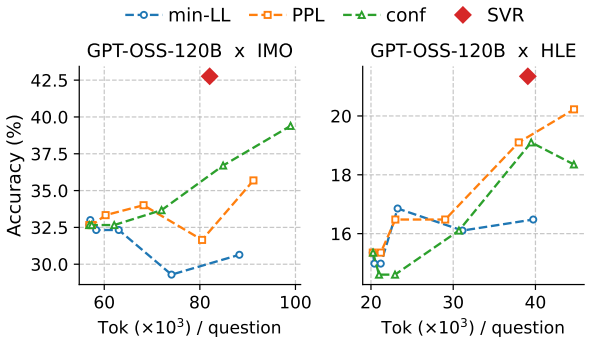

Figure 3: **SVR yields better token–accuracy tradeoffs under a fixed communication control flow.** Token–accuracy tradeoffs when replacing SVR with alternative posterior signals, at acceptance rates from 10% to 90%.

sures, independent of Algorithm 1's control flow. To isolate this effect, we keep the algorithm unchanged and replace only the posterior score computation on line 9. For agent $\mathcal{A}_r$, let D denote the number of debates involving $\mathcal{A}_r$ as receiver, and let $\mathcal{P}$ denote the corresponding peer agents. For each alternative signal, we compute the posterior score: $Po(\mathcal{A}_r) = \frac{1}{D}\sum_{\mathcal{A}_s \in \mathcal{P}} Signal(\text{Debate}(\mathcal{A}_s, \mathcal{A}_r))$ where $Signal$ is min. LL, PPL, or self-confidence extracted from $\mathcal{A}_r$'s post-debate responses.

Figure 3 shows the token–accuracy tradeoff on IMO and HLE using GPT-OSS. For each alternative signal, we sweep the acceptance threshold to cover acceptance rates from 10% to 90%, and compare against the default Algorithm 1 using SVR. Overall, SVR provides a better token–accuracy tradeoff than alternatives. At matched token cost, SVR improves accuracy by +8.75 and +2.25 percentage points over the best alternative. These results demonstrate SVR's effectiveness independent of the communication control-flow.

5.2.2 Validation of the SVR assumption

SVR-MAD relies on the assumption that correct agents are more likely to retain their answers after being challenged. We further empirically validate this assumption on the same setting as Figure 1. Among agents that revise their answer after at least one challenge (SVR<1), only 14.5% and 15.4% are correct for GPT-OSS and DeepSeek. In contrast, agents that retain their answer after every challenge (SVR=1) are correct 70.3% and 57.1% of the time, 4.8× and 3.7× higher than agents with $SVR < 1$. This supports SVR as a strong empirical signal of correctness. However, the SVR assumption may not hold on smaller LLMs with limited reasoning capabilities: those models may be easily persuaded by any peer arguments that appear to be sound. Appendix C shows one failure case with the Qwen3 family, where SVR-MAD is effective on the 235B but collapses on the 32B variant.

N	All-to-all	GPT-OSS		DeepSeek	
		IMO	HLE	IMO	HLE
3	6	2.6	2.8	2.9	2.0
4	12	3.8	4.5	5.9	3.6
5	20	4.8	6.1	7.3	4.6
6	30	5.6	7.3	8.9	5.3

Table 2: **SVR-MAD scales efficiently with the number of agents.** Average NComm. as the number of agents N increases. The All-to-all column reports the cost of unpruned all-to-all communication. The remaining columns report SVR-MAD per <dataset, LLM> setting.

5.2.3 Sensitivity to prior signals

SVR-MAD uses a single type of pre-debate signal to initialize correctness scores, with PPL for GPT-OSS and min-LL for DeepSeek-V3.1. Since SVR-MAD's outcome is largely determined by posterior evidence, the initialization has a small effect. We evaluate the sensitivity of prior signal choice by evaluating SVR-MAD with all three discussed prior signals from §3. Accuracies under different prior choices do not shift significantly (paired McNemar $p > 0.05$ for all <dataset, LLM> cases), while NComm and tokens move by less than 10% in any setting, well within the 48–75% and 38–61% reduction range from Table 1.

5.2.4 Scaling with the number of agents

Table 2 reports the NComm of SVR-MAD as the number of agents increases. While the all-to-all grows by 5× from $N = 3$ to 6, SVR-MAD grows by only 2.2×–3.1× across the four settings. Overall, SVR-MAD's relative communication savings increase as the number of agents grows.

6 Conclusion

We present SVR-MAD, a Bayesian-inspired MAD framework that leverages posterior evidence from debate outcomes. SVR-MAD probes selected communications to estimate agent reliability, then incrementally constructs the communication graph toward agents with stronger posterior signals. Evaluation shows SVR-MAD improves the cost-accuracy trade-off, reducing communication links and tokens by up to 75% and 61%, while matching or improving accuracy relative to the most accurate competing MAD baseline.

Limitations

We discuss several limitations of this work. First, SVR is an evidence-grounded proxy for correctness. Our central assumption is that agents with sound, well-supported reasoning are more likely to retain their answers when challenged by peers. While our results show that SVR is more robust than prior signals on hard problems, it does not guarantee correctness. Incorrect agents may also retain their answers when peer arguments are weak or when the model is overconfident in an incorrect reasoning path. A higher acceptance threshold can partially mitigate this issue, and we plan to further improve SVR reliability in future work.

Second, our evaluation focuses on closed-ended reasoning tasks where final answers can be compared automatically. This setting enables consistent majority voting and correctness measurement, but limits the generality of our claims. In particular, it remains unclear whether SVR-MAD would provide similarly reliable signals for tasks such as open-ended generation, factual QA, code generation, or planning, where answer equivalence and correctness are harder to determine automatically. Extending SVR-MAD to such settings may require alternative mechanisms for measuring answer agreement and evaluating debate outcomes, which we leave to future work.

Third, the effectiveness of SVR-MAD may depend on the debate prompt, the number of agents, and the model family. Different LLMs may respond differently to peer challenges: some may revise too aggressively, while others may retain answers even when presented with strong counterarguments. We plan to further improve SVR-MAD's robustness by studying model-specific debate prompts and validating the method across more model families and domains.

Ethical considerations

This work does not raise any ethical concerns.

Acknowledgments

We thank the reviewers and the AC for their service and constructive feedback. Michael Mitzenmacher and Minlan Yu are supported by NSF CNS-2107078: CNS Core: Medium: Approximation and Randomization in the Programmable Data Plane. This work was supported in part by ACE, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

Xuhang Chen, Zhifan Song, Deyi Ji, Shuo Gao, and Lanyun Zhu. 2026. Self-signals driven multi-llm debate for efficient and accurate reasoning. *Preprint*, arXiv:2510.06843.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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

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

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *Preprint*, arXiv:2509.04664.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2026. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '26, page 1147–1155, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Thang Luong, Dawsen Hwang, Hoang H Nguyen, Golnaz Ghiasi, Yuri Chervonyi, Insuk Seo, Junsu Kim, Garrett Bingham, Jonathan Lee, Swaroop Mishra, Alex Zhai, Huiyi Hu, Henryk Michalewski, Jimin Kim, Jeonghyun Ahn, Junhwi Bae, Xingyou Song,

Trieu Hoang Trinh, Quoc V Le, and Junehyuk Jung. 2025. Towards robust mathematical reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35418–35442, Suzhou, China. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 281 others. 2026. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649(8099):1139–1146.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and Xiaohua Xu. 2025. S^2-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9393–9408, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. 2026. Demystifying multi-agent debate: The role of confidence and diversity. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 33909–33930, San Diego, California, United States. Association for Computational Linguistics.

A Additional Experimental Settings

A.1 Dataset pre-processing

As described in §5, we filter HLE to text-only, non-math, multiple-choice questions. For each <LLM, dataset> setting, we remove questions on which all six agents give the same answer before debate, since debate is unnecessary in such cases. Note that we do keep questions where all agents are wrong but disagree on the answer.

	IMO	HLE
GPT-OSS-120B	297	267
DeepSeek-V3.1	299	276

Table 3: Valid sample size per experimental setting after all pre-processing steps.

The questions removed from each <LLM, dataset> setting differ, so Table 1 is not suitable for cross-setting comparisons of the same MAD method. However, the same set of questions is used for all MAD methods under each <LLM, dataset>. Table 3 records the valid sample size per evaluation setting.

A.2 Generation and baseline configurations

We keep default sampling parameters for all settings: temperature 1.0, top-p 0.95, top-k 40. We set the maximum output tokens to 16,384 per LLM response. We use medium reasoning effort for both LLMs, as higher efforts often produce long reasoning chains that exhaust the maximum token cap before outputting a valid answer.

GroupDebate. For the 6-agent setting, we consider two grouping configurations: two groups of three (3,3) and three groups of two (2,2,2). We evaluated both settings. As shown in Table 4, both grouping configurations trail SVR-MAD in accuracy and use more tokens. We use (3,3) in the main evaluation results shown in Table 1 for its higher accuracy.

S^2-MAD. S^2-MAD proposes two approaches to identify agents with similar or identical viewpoints, which guides communication pruning decisions: (1) answer equivalence for tasks with a closed-form answer (e.g., math), and (2) response embedding similarity when answers are not directly comparable for equivalence (e.g., coding). We use (1) for both IMO-Answerbench (math) and HLE (multiple-choice) benchmarks.

SID-ET. SID-ET applies a threshold over a single agent's pre-debate min. LL, to determine if the question is solvable by a single agent, or requires all-to-all MAD. We sweep threshold values that lead to different skip rates (i.e., X% questions are skipped from MAD due to sufficient single-agent min. LL, for X=$\{10, 20, ..., 90\}$). A higher skip

Method	NComm↓	Tok($\times 10^3$) ↓	Acc↑
GPT-OSS-120B / IMO			
GroupDebate (3,3)	19.4	164.1	41.8
GroupDebate (2,2,2)	17.1	150.7	40.4
SVR-MAD	**5.6**	**82.1**	**42.8**
GPT-OSS-120B / HLE			
GroupDebate (3,3)	18.4	111.0	20.2
GroupDebate (2,2,2)	16.0	76.6	19.1
SVR-MAD	**7.3**	**39.1**	**21.4**
DeepSeek-V3.1 / IMO			
GroupDebate (3,3)	19.8	234.6	40.1
GroupDebate (2,2,2)	17.4	178.4	37.1
SVR-MAD	**8.9**	**91.7**	**41.5**
DeepSeek-V3.1 / HLE			
GroupDebate (3,3)	19.6	108.7	14.9
GroupDebate (2,2,2)	18.9	100.4	14.5
SVR-MAD	**5.3**	**30.3**	**16.7**

Table 4: **SVR-MAD achieves higher accuracy and lower communication cost than both GroupDebate configurations.** Accuracy and communication costs for SVR-MAD and GroupDebate with two grouping configurations.

rate reduces communication costs but may also reduce accuracy by aggressively skipping debates.

Algorithm 2 Skip rate tuning for SID-ET.

Input: (1) Tok_{ref}: Total token used by SVR-MAD, (2) Acc_{ref}: Accuracy of SVR-MAD.
Output: Selected skip rate $\hat{r}$

Initialize skip rate $r = 90\%$
while $r \geq 10\%$ **do**
  $Tok_{SID}, Acc_{SID} = SID - ET(r)$
  if $Acc_{SID} \geq Acc_{ref}$ **then**
    return $\hat{r} = r$
  else if $Tok_{SID} > Tok_{ref}$ **then**
    return $\hat{r} = r$
  end if
  $r \mathrel{-}= 10\%$
end while
return $\hat{r} = 10\%$

We tune SID-ET's skip rate per <LLM, dataset> setting. We aim to find an optimal skip rate that Pareto-dominates SVR-MAD in terms of tokens and accuracy. If not found, we report results using the SID-ET setting that just exceeds the token cost of SVR-MAD. This tuning procedure is documented in Algorithm 2. Overall, no setting of SID-ET Pareto-dominates SVR-MAD in terms of tokens and accuracy. The final selected skip rates are 60/70 for IMO-AnswerBench/HLE using GPT-OSS-120B, and 50/60 using DeepSeek-V3.1.

Note that the original SID paper (Chen et al., 2026) has a second technique, which compresses message size in all-to-all MAD based on token-level attention scores. This compression technique is triggered when early-stopping fails, and it only reduces token cost but not the number of communications. Our experiments use larger LLMs than (Chen et al., 2026) (i.e., hundreds vs. tens of billions of parameters), making this attention-driven compression computationally expensive to evaluate. However, SVR-MAD still saves on the number of cross-agent communications.

A.3 SVR-MAD configuration

Sender selection criteria: SVR-MAD supports two ranking criteria when selecting the most challenging senders for a receiver (line 7, Algorithm 1): top-S by $Corr$ or by $Prior$. $Corr$ prioritizes senders whose arguments are validated by post-debate evidence, while $Prior$ favors those appearing most sound by pre-debate signals. We use $Prior$ in all evaluations. However, **this choice does not affect algorithm performance in our experiments:** switching the sender ranking to $Corr$ moves SVR-MAD's accuracy by -0.34 to $+0.36$ percentage points across the four settings of Table 1, and NComm and tokens by -0.1 to $+0.3\%$ and -0.3 to $+0.4\%$, respectively. None of these shifts are statistically significant (paired McNemar, $p = 1.0$). At six agents per question, the two rankings often select the same senders. An extended analysis with more agents per question would further separate the two criteria, which we leave as future work.

Acceptance threshold (τ): In our implementation, we accept an answer if the agent achieves 100% SVR against at least C disagreeing peers (or against all disagreeing peers if it has fewer than C). We set $C = 2$ for GPT-OSS-120B and 3 for DeepSeek-V3.1. This is because agents in DeepSeek-V3.1 typically disagree more before debate; as shown in Table 3, for each dataset, using DeepSeek-V3.1 yields more non-unanimous pre-debate questions. Therefore, we set a higher acceptance threshold for DeepSeek-V3.1 agents to ensure disagreements are sufficiently evaluated.

Note that for a strict acceptance condition of 100% SVR, a single failed survival makes acceptance impossible for the agent. Under this condition, our implementation excludes agents with failed challenges from receiver selection, avoiding communications that can no longer lead to accep-

tance.

Communications per round (S): We set $S = 2$ for GPT-OSS-120B, and $S = 3$ for DeepSeek-V3.1, based on their respective acceptance thresholds.

Communication budget per problem (B_{max}): We set $B_{\max} = S \cdot (k + m)$, where k is the number of distinct pre-debate answer clusters and m is the size of the largest cluster. The first term provides budget for one probed representative per cluster; the second adds budget to cover all members of the largest cluster. This simple rule works well in our experiments; a more principled budget allocation that further tightens costs is an interesting direction for future work.

A.4 Prompts

MAD performance is sensitive to prompting. In this work, we apply fixed prompting instructions to all baselines and vary only the communication control flow for a fair comparison. All prompts used in our experiments are in Figure 4.

B Statistical uncertainty of results

We assess the statistical uncertainty of the accuracy results in Table 1 using paired McNemar tests for each SVR-MAD-against-baseline comparison and 10,000-sample paired bootstrap confidence intervals for accuracy differences. Pooled across the four <dataset, LLM> settings, SVR-MAD improves over SID-ET by +4.7 percentage points ($p = 1.1 \times 10^{-5}$). Against S2-MAD, the improvement is significant on DeepSeek-V3.1/IMO (+4.7 points, $p = 0.044$); the remaining settings show no statistically significant accuracy difference, while SVR-MAD uses 38–58% fewer tokens. Against GroupDebate, SVR-MAD shows no significant accuracy difference ($p = 0.41$–0.71), with a pooled bootstrap 95% CI of $[-0.4, +3.1]$ points (favoring SVR-MAD), while using 50–72% fewer tokens.

Overall, the statistical analysis supports the pattern in Table 1: SVR-MAD either significantly improves accuracy or shows no significant accuracy difference while substantially reducing inference cost.

C Preliminary evaluation on additional LLM families

We provide preliminary evaluation results on two additional LLMs: Qwen3-32B and Qwen3-235B (Yang et al., 2025), on IMO-AnswerBench using 4 agents per question. We compare SVR-MAD to the no-debate baseline, Self Consistency. SVR-MAD uses min-LL as the prior signal and accepts an answer if it survives at least 3 challenges.

As shown in Table 5, SVR-MAD improves accuracy by 5.6 percentage points over Self Consistency for the large-scale Qwen3-235B backbone, comparable to the margin of improvement obtained on GPT-OSS. However, SVR-MAD trails Self Consistency in accuracy by 1.6 percentage points on Qwen3-32B, the smaller-scale backbone within the same LLM family. This demonstrates the failure mode discussed in §5.2.2, in which smaller models may have limited instruction-following capabilities and consequently violate the SVR assumption.

D Effect of greedy peer ordering

Algorithm 1 probes S communication links per agent, greedily selecting top-S disagreeing peers with the highest $Corr$ or $Prior$ scores. Our experiments use $Prior$ as the criterion. This section analyzes the magnitude of the greedy selection bias and whether it affects algorithm outcomes.

We compute the unbiased SVR for each agent from all of its disagreeing peers. Overall, the unbiased SVR values do not change the accept/reject decisions for 94-97% of agents across four <dataset, LLM> settings, demonstrating that SVR-MAD's outcomes are robust to greedy estimation bias. In magnitude, the mean difference between the greedy and unbiased SVR ranges from -0.05 to $+0.01$ across four settings, small on a scale where SVR ranges from -1 to 1. Greedy estimates are smaller than unbiased values for 77–87% of agents, as SVR-MAD prioritizes the strongest challengers and avoids accepting an agent prematurely.

E Use of AI Assistants

We used AI assistants for writing polish and code development support. All research ideas, methodological decisions, experimental design, analysis, and conclusions are the authors' own work.

LLM	Method	NComm↓	Tok($\times 10^3$) ↓	Acc↑
Qwen3-235B	Self Consistency	0.00	39.7	28.93
	SVR-MAD	6.64	187.4	**34.52**
Qwen3-32B	Self Consistency	0.00	10.6	**16.99**
	SVR-MAD	6.04	48.9	15.34

Table 5: **SVR-MAD transfers to Qwen model family but requires a sufficiently large backbone.** Accuracies and communication costs on IMO-AnswerBench for SVR-MAD and Self Consistency, using two LLM backbones in the Qwen3 family.

(a) Initial response generation, IMO-AnswerBench

```
You are an expert mathematician. Solve the following problem step by step, showing your complete
reasoning.

Problem: {Question}

Your response must end with:
\boxed{<your final answer>}
Confidence: <0-100%>
```

(b) Initial response generation, HLE

```
You are an expert problem solver. The following is a multiple-choice question. Read it carefully,
reason step by step, then choose the single best answer.

Question: {Question}

Your response must end with:
\boxed{<letter of your chosen answer, e.g. A>}
Confidence: <0-100%>
```

(c) Debate prompt, GPT-OSS-120B, IMO

```
These are the solutions to the problem from other agents:
{peer_content}
Using their solutions as additional information, provide your answer to the problem. Your response
must end with your final answer in \boxed{<answer>} followed by Confidence: <0-100%>.
```

(d) Debate prompt, GPT-OSS-120B, HLE

```
These are the answers and reasoning from other agents:
{peer_content}
Using their answers as additional information, choose the single best answer. Your response must end
with the letter of your chosen answer in \boxed{<letter>} followed by Confidence: <0-100%>.
```

(e) Debate prompt, DeepSeek-V3.1, IMO

```
These are the solutions to the problem from other agents:
{peer_content}
Critically evaluate their reasoning. If your prior reasoning is mathematically sound, defend it. Only
 revise your answer if a peer's argument identifies a genuine error in yours. Your response must end
with your final answer in \boxed{<answer>} followed by Confidence: <0-100%>.
```

(f) Debate prompt, DeepSeek-V3.1, HLE

```
These are the answers and reasoning from other agents:
{peer_content}
Critically evaluate their reasoning. If your prior reasoning is sound, defend it. Only revise your
answer if a peer's argument identifies a genuine error in yours. Your response must end with your
final answer in \boxed{<answer>} followed by Confidence: <0-100%>.
```

Figure 4: **Prompts used in all experiments.** (a–b) generate each agent's initial response for the two tasks. (c–f) instruct the agent to reflect on peer response and update their own answers if needed.