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
https://github.com/weifanjiang/SVR-MAD

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., 2025; Chen et al., 2025; 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 min. log-likelihood values are often considered to be reliable, and therefore are likely to be correct without further debate (Chen et al., 2025).

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 propose *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 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, each agent's context grows as $\mathcal{O}(N^2t)$, increasing computation and KV-cache pressure while also risking long-context degradation from diluted salient information (Liu et al., 2024; Laban et al., 2026).

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 can therefore conclude without any debate (Chen et al., 2025; Eo et al., 2025). Another line uses cross-agent similarity signals to remove communication links that appear redundant (Zeng et al., 2025; Liu 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: min. 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).

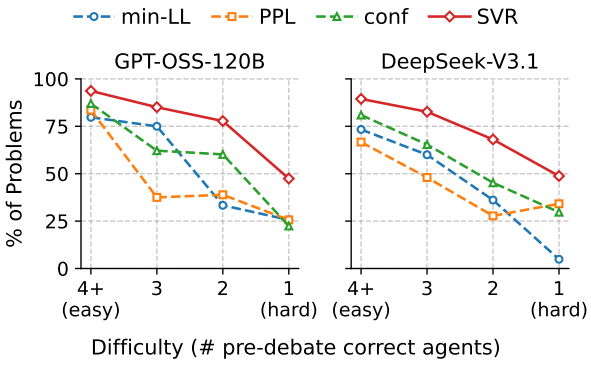

Figure 1: **SVR better identifies correct agents than prior signals under increasing difficulty.** Percentage of problems where the top-ranked agent under each signal is correct, as a function of increasing difficulty.

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, 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, 2024), using questions from IMO-AnswerBench (Luong et al., 2025). For each question, we run six agents and group questions by difficulty, measured by the number of agents that are correct before debate. We compare three prior signals against SVR by measuring how often the top-ranked agent under each signal is correct.

Results. Figure 1 reports this percentage across difficulty groups, showing how each signal aligns with agent correctness as problems become harder. From 4+ correct agents (easiest) to 1 correct agent (hardest), the correctness of top-ranked agents under min-LL, PPL, and confidence drops from 67–87% to 5–34%. **This sharp decline shows that prior signals become unreliable on difficult problems, where hallucinated or overconfident reasoning makes model behavior poorly aligned with correctness.** In contrast, SVR remains predictive as difficulty increases: among questions with only 2 and 1 correct agents, SVR achieves 68–78% and 47–49%, respectively, outperforming the strongest prior signal by 18–23 and 15–22 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 $Corr[\mathcal{A}_i] = \text{Prior}(\mathcal{A}_i)$ for $1 \le i \le N$
2: **Initialize** $B = B_{max}$
3: **while** $B > 0$ **do**
4: $\quad \mathcal{A}_r = \arg\max_{\mathcal{A}} Corr[\mathcal{A}]$
5: $\quad$ Disagree $= \{\mathcal{A}_j : y_j \neq y_r\}$
6: $\quad \mathcal{P} =$ top-S of Disagree ranked by $Corr$
7: $\quad$ **for all** $\mathcal{A}_s \in \mathcal{P}$ **do**
8: $\quad\quad Outcome = Debate(\mathcal{A}_s \to \mathcal{A}_r)$
9: $\quad\quad Corr[\mathcal{A}_r] \mathrel{+}= \ell(Outcome)$
10: $\quad$ **end for**
11: $\quad$ **if** $Corr[\mathcal{A}_r] \ge \tau$ **then**
12: $\quad\quad$ **return** $\hat{y} = y_r$
13: $\quad$ **end if**
14: $\quad B \mathrel{-}= S$
15: **end while**
16: **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 pre-debate prior signals, which provide an initial estimate of whether A_i holds a correct answer. These scores are updated as debate outcomes become available.

Incremental graph construction (line 3-15): 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. 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 4-6): In each iteration, SVR-MAD selects the agent with the highest correctness score as the receiver. This greedy strategy prioritizes agents most likely to hold correct answers, helping SVR-MAD confirm reliable solutions and terminate with fewer communications. Among peers that disagree with the receiver $\mathcal{A}_r$, SVR-MAD selects up to S strongest challengers based on their correctness scores.

Posterior-style score update (line 7-9): SVR-MAD updates each receiver's correctness score after every debate. For receiver $\mathcal{A}_r$, 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 $c \mathrel{+}= 1$ based on debate outcome. We then apply a posterior-dominant rule:

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

Thus, prior signals guide selection only before debate evidence is available; once observed, SVR dominates and is refreshed after each debate.

Early termination (line 11-13): After each debate, SVR-MAD checks whether the receiver's updated correctness score exceeds the acceptance threshold. If so, SVR-MAD terminates the debate early and returns that agent's answer as the final solution to reduce overall communication cost.

Fallback majority-voting (line 16): If no agent commits within budget, each agent contributes one vote: the majority of its observed post-debate answers, with its pre-debate answer y_i as tiebreaker. SVR-MAD returns the majority over the N votes, tie-breaking by the pre-debate majority among $y_1, \dots, y_N$.

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., 2025), which restricts full-context exchange to random subgroups and allows only final-answer messages across groups; (3) SID-ET (Chen et al., 2025), which uses token log-likelihood-based self-confidence to select agents for debate; and (4) S^2-MAD (Zeng et al., 2025), which prunes communication links between agents with similar answers or response embeddings. 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

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).

may terminate after round one if at least five agents reach a clear consensus, thereby avoiding unnecessary cost inflation.

LLMs and datasets. We evaluate on GPT-OSS-120B and DeepSeek-V3.1 using IMO-AnswerBench and a filtered HLE (Phan et al., 2026) subset containing only text-only, non-math, multiple-choice questions. This filtering ensures compatibility with GPT-OSS-120B (text-only LLM), avoids domain overlap with IMO-AnswerBench, and enables automatic correctness evaluation. We skip questions where all six agents share a unanimous answer before debate.

Additional experimental settings. We include other experimental details 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 points compared to GroupDebate, SID, and S^2-MAD, respectively. On DeepSeek-V3.1/HLE, SVR-MAD matches S^2-MAD's accuracy but uses 72%/58% fewer communications/tokens per question.

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 GPT-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.

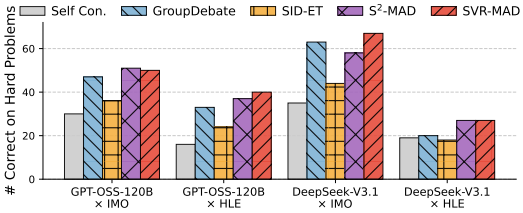

Figure 2: **SVR-MAD achieves high accuracy on hard problems across all evaluated settings.** Accuracy of each method on problems where the number of correct agents before debate is within 1-3.

Ablation studies. Appendix B isolates the contribution of SVR by replacing it with alternative posterior signals in Algorithm 1. SVR improves accuracy by up to 8.75 points at matched token cost compared to the best alternative, showing that SVR provides an effective posterior signal beyond the benefits of the communication control flow.

6 Conclusion

We present SVR-MAD, a Bayesian-inspired MAD framework that leverages posterior evidence from debate outcomes. SVR-MAD first probes selected communications to estimate agent reliability, then incrementally constructs the communication graph toward agents with stronger posterior signals. Evaluation across multiple LLMs and datasets 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 it does not fully capture open-ended generation tasks, where determining answer equivalence is harder. We plan to extend SVR-MAD to open-ended tasks in 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.

References

Xuhang Chen, Zhifan Song, Deyi Ji, Shuo Gao, and Lanyun Zhu. 2025. Sid: Multi-llm debate driven by self signals. *Preprint*, arXiv:2510.06843.

DeepSeek-AI. 2024. 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.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *The Fourteenth International Conference on Learning Representations*.

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. 2025. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *Preprint*, arXiv:2409.14051.

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*.

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. *Preprint*, arXiv:2601.19921.

A Additional Experimental Settings

We provide additional experimental details to ensure full reproducibility of our work.

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 2: 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 2 records the valid sample size per evaluation setting.

LLM settings. 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 set the reasoning effort to medium for both LLMs, as higher efforts often produce long reasoning chains that exhaust the maximum token cap before outputting a valid answer.

GroupDebate. We consider two possible settings: two groups of three (3-3), and three groups of two (2-2-2). The original paper suggests that (3-3) yields higher accuracy while (2-2-2) uses fewer tokens (Table 2 in (Liu et al., 2025)). Therefore, we tested (3-3) first, and planned to switch to (2-2-2) if (3-3) yields better accuracy than SVR-MAD. The GroupDebate results in Table 1 are all using (3-3). We did not further evaluate (2-2-2), as (3-3) already trails SVR-MAD in accuracy among all <LLM, dataset> settings.

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 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 doc-

umented in Algorithm 2. Overall, no settings of SID-ET Pareto-dominates SVR-MAD in terms of tokens and accuracy. The final selected skip rates are in 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., 2025) 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., 2025) (i.e., hundreds vs. tens billions of parameters), making this attention-driven compression computationally expensive to evaluate. However, SVR-MAD still saves in the number of cross-agent communications.

SVR-MAD. We detail the configuration scheme for key hyperparameters in SVR-MAD.

- Acceptance threshold (τ): In our implementation, we accept an answer if the agent achieves 100% SVR for at least C challengers. We set $C = 2$ for GPT-OSS-120B and $C = 3$ for DeepSeek-V3.1. This is because agents in DeepSeek-V3.1 typically have more disagreements before debate: as shown by Table 2, 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.
- 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.

Additional implementation details. Additional implementation decisions such as prompts used for initial response generation and MAD facilitation are in provided code implementations.

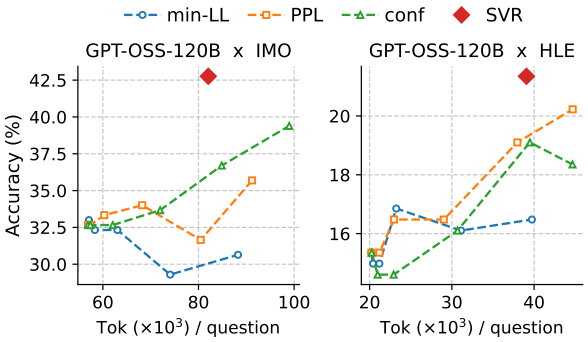

Figure 3: **Ablation shows that SVR improves posterior scoring.** Token–accuracy tradeoffs when replacing SVR with alternative post-debate posterior signals, evaluated across acceptance rates from 10% to 90%.

B Ablation Studies

We isolate the contribution of SVR by replacing it in Algorithm 1 while keeping the same communication-control procedure. Specifically, we modify 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 set of peer agents. We compute:

$$\text{Posterior}(\mathcal{A}_r) = \frac{1}{D} \sum_{\mathcal{A}_s \in \mathcal{P}} \text{Signal}(\text{Debate}(\mathcal{A}_s, \mathcal{A}_r)) \,.$$

Here, Signal is one of min. LL, PPL, or self-reported confidence, extracted from $\mathcal{A}_r$'s post-debate responses. Thus, instead of using SVR to measure whether $\mathcal{A}_r$ retains its answer after debate, these ablations use post-debate confidence-like signals averaged across all debates.

Figure 3 shows the token–accuracy tradeoff on IMO-AnswerBench and HLE using GPT-OSS-120B. We compare the default SVR-MAD, which uses SVR as the posterior score, against variants that replace SVR with alternative post-debate signals. For each variant, we sweep the acceptance threshold to cover acceptance rates from 10% to 90%.

Overall, SVR demonstrates more optimal token–accuracy tradeoffs than alternative posterior score choices. SVR achieves higher accuracy than alternatives that fail to match it under all tested acceptance rates, with +3.37 and +1.13 points on IMO and on HLE compared to the strongest competitor. At matched token cost, SVR improves accuracy by +8.75 and +2.25 points compared to the best alternative signal on the two datasets, respectively. This shows that SVR is an effective posterior signal

independent of the communication control-flow in Algorithm 1.

C 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.