ASMAD: Adaptive Sparse Communication Topology Multi-Agent Debate Framework with Opinion Dynamics

Anonymous ACL submission

Abstract

Large language models (LLMs) still face challenges in complex reasoning within multi-agent debate (MAD) systems due to high computational costs in fully-connected structures. While existing methods use static sparse topologies to reduce computation, they neglect semantic relationships and dynamic opinion evolution. To solve this challenge, we propose ASMAD, an adaptive sparse topology framework that synergizes sociophysical opinion dynamics with LLMs through two innovations: (1) probabilistic semantic-guided attention gates for dynamic opinion visibility control; (2) a hybrid paradigm combining adaptive trust-boundary regulation and opinion synchronization. Experiments show ASMAD reduces token costs to around 1/3 across GSM8K and MMLU benchmarks while maintaining competitive accuracy with 4-bit quantized 7-9B size models.

1 Introduction

In recent years, the rapid development of large language models (LLM) has greatly promoted the progress of several natural language processing (NLP) tasks (Touvron et al., 2023; Zhao et al., 2023; Naveed et al., 2023; Jiang et al., 2024; Achiam et al., 2023; GLM et al., 2024; Guo et al., 2025). However, performance of LLM in reasoning and logical reasoning tasks is still limited (Zhu et al., 2022; Gou et al., 2023).

To address complex reasoning challenges, various approaches has been developed, including Chain-of-Thought (CoT) (Wei et al., 2022), self-consistency (SC) mechanisms (Wang et al., 2022) with self-correction strategies (Liang et al., 2023). Recent advances in multi-agent debate (MAD) systems have demonstrated superior performance in complex reasoning tasks (Liang et al., 2023). Inspired by the human discussion mechanism (Hill et al., 2015; Liang et al., 2023), MAD systems employ multiple LLM agents to communicate and iteratively argue with each other in a structured debate. However, MAD systems face computation cost problem due to fully-connected communication topology, where every agent interacts with all peers, which incurs quadratic computational complexity that becomes prohibitively expensive for real-world applications (Du et al., 2023).

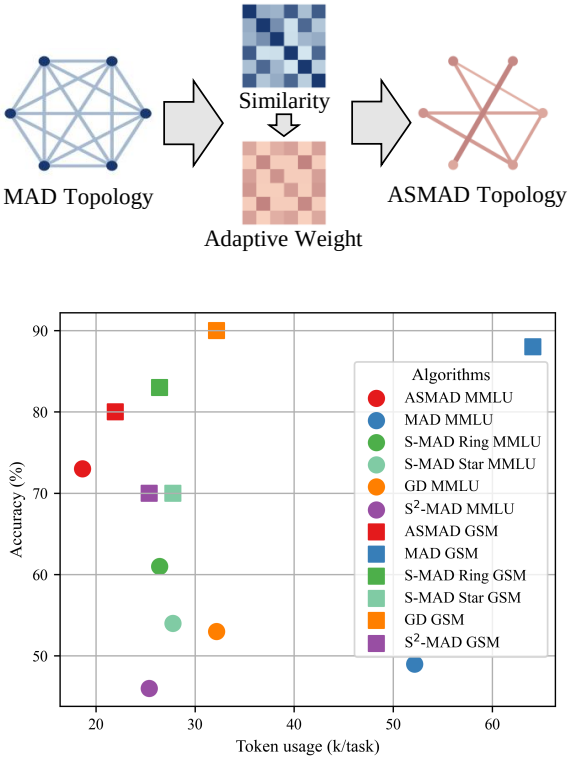

Figure 1: Adaptive topology of ASMAD (Top) and comparison of accuracy and token consumption (Bottom). The results show that we achieve better cost-acc trade.

Existing attempts to address this efficiency challenge focus on either static sparse topologies (e.g., ring or star structures of Sparse MAD (S-MAD) (Li et al., 2024)) that reduce token costs through predetermined connection patterns (Du et al., 2023; Sun et al., 2023) or group discussion methods like Group Debate (GD) (Liu et al., 2024) or Selective Sparse MAD (S^2-MAD) (Zeng et al., 2025) that

adopts a hierarchical structure by clustering agents into smaller debate groups to exchange intermediate results. However, existing approaches face two fundamental limitations: (1) Task-semantic blindness: fixed topologies cannot adapt to problem difficulty, potentially pruning critical debate pathways; (2) Coarse adaptation granularity: fixed grouping patterns cannot capture nuanced opinion evolution dynamics.

To address these limitations, we propose a adaptive sparse topology framework (ASMAD) that synergies sociophysical opinion dynamics with modern LLM architectures as shown in Figure 1. Our key insight stems from two observations: First, human consensus formation naturally evolves communication networks through confidence-bound adaptation, suggesting that artificial debate systems should similarly adjust interaction patterns based on semantic convergence states. Second, semantic similarity between textual opinions provides a more reliable signal for trust boundary calculation than numerical difference metrics.

Building upon this foundation, we propose a dual-regulation debate mechanism that hybridizes two classical models: The Hegselmann-Krause (HK) model (Rainer and Krause, 2002) inspired adaptive trust boundary allows agents to dynamically adjust their openness to divergent views based on real-time semantic proximity, while the Deffuant model (Deffuant et al., 2000) derived synchronization protocol coordinates opinion aggregation through gradient descent in the semantic space. The system's core innovation lies in visibility control module, which implements selective opinion exposure through attention-based gates.

We evaluate ASMAD across GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021) benchmarks[1] using 4-bit quantized versions of LLaMA-8B (Touvron et al., 2023), ChatGLM-9B (GLM et al., 2024) and Deepseek-7B (Guo et al., 2025). Experiments show ASMAD reduces token costs up to 65.8% while maintaining competitive accuracies. Figure 1 shows ASMAD gets better cost-accuracy trade-off.

In summary, our work contributes as following:

- We developed dynamic visibility control mechanisms for agent opinions in MAD with lower cost and better consensus.
- We extended classical opinion dynamics models to LLM-based MAD systems through a tunable debate paradigm integrating Deffuant model's adaptive trust-boundary regulation with HK model's synchronized opinion aggregation.
- We introduced a methodology replacing conventional numerical handcrafted metrics with SentenceTransformer-based semantic vectors and similarity matrices. It might be a potential workaround for LLM multi-agent systems to effectively handle unstructured textual opinions.

[1]MIT License

2 Related Works

Topology in MAD Due to the diversity of human discussion strategies (Liang et al., 2023; Chan et al., 2023; Du et al., 2023), researchers adjust the visibility of interactions between agents and their historical records as well as among the agents themselves, by employing different multi-agent topologies, ultimately reducing token cost or enabling operation in resource-constrained environments (Li et al., 2024; Liu et al., 2024).

Regarding historical records, Du et al. (2023) process information from a centralized topology by summarizing agent outputs at the end of each round, whereas Sun et al. (2023) introduces a forgetting mechanism in which agents can only see the outputs from the previous round. In addition, Zhang et al. (2023) proposes a debate–reflection mechanism in which agents can only review their own past outputs during reflection.

Several studies focus on the topology of inter-agent information exchange. For instance, S-MAD (Li et al., 2024) employs a sparse topology, limiting information exchange to adjacent agents. GroupDebate (GD) (Liu et al., 2024) adopts a hierarchical structure by clustering agents into smaller debate groups to exchange intermediate results. Furthermore, S^2-MAD (Zeng et al., 2025) utilizes a sparse topology based on grouping and a decision mechanism: agents initially generate independent opinions within groups, and only engage in information exchange within and between groups if a decision mechanism identifies differences in opinions.

Opinion Dynamics In the study of opinion dynamics, the Deffuant model and Hegselmann-Krause (HK) dynamics (Deffuant et al., 2000; Rainer and Krause, 2002) serve as foundational consensus models where a group of agents strive to reach the same objective. The Deffuant model posits that agents update their opinions based on a

bounded confidence mechanism: two agents adjust their opinions only when their difference falls below a predefined threshold (Deffuant et al., 2000, 2002; Lorenz, 2007). This model has been extensively applied to investigate opinion convergence and polarization phenomena in social networks (Zhang et al., 2017; Marconi and Cecconi, 2020; Zarei et al., 2023).

The Hegselmann-Krause (HK) dynamics assumes that agents interact exclusively with peers whose opinions lie within their confidence bounds (Rainer and Krause, 2002; Etesami and Başar, 2015). In its synchronous variant, agents simultaneously update opinions by averaging those of neighbors within their confidence interval (Rainer and Krause, 2002; Etesami et al., 2013; Etesami and Başar, 2015), whereas the asynchronous version updates one agent at a time (Rainer and Krause, 2002; Touri and Langbort, 2014; Etesami and Başar, 2015). These consensus models provide critical frameworks for understanding opinion formation and evolution in social systems, particularly in analyzing how local interactions drive collective behaviors.

3 Methodology

3.1 Dynamic Opinion Exchange Framework

Multi-agent debate (MAD) with large language models presents unique challenges that traditional frameworks struggle to address. This work reframes the MAD process through the theoretical lens of opinion dynamics, treating each LLM as an agent with bounded rationality, whose willingness to incorporate external viewpoints varies dynamically based on semantic proximity and confidence levels. Drawing from both HK and Deffuant models, we implement: **Simultaneous Updates**: All agents update their states based on visible information, **Probabilistic Interaction**: Probabilities and strength of pairwise interaction determined by adaptive weights.

Unlike classical opinion dynamics that operate in numerical spaces, our framework extends into rich semantic embeddings where agent states comprise both reasoning processes and discrete conclusions. We introduce the agent state as $s_i^t = (r_i^t, c_i^t)$, where $r_i^t \in \mathbb{R}^d$ represents the semantic embedding of agent i's reasoning at time t, and c_i^t denotes its conclusion. This richer state space enables more nuanced modeling of debate dynamics while preserving the mathematical tractability of opinion evolution.

3.2 Adaptive Debate Protocol

As detailed in Figure 2, the proposed protocol orchestrates multi-agent debate through distinct phases that progressively refine agent opinions while maintaining diversity and efficiency.

Independent Initialization Each agent independently generates its initial response to the given problem without access to other agents' outputs. Formally, at $t = 0$, agent i produces state $s_i^0 = (r_i^0, c_i^0)$, where r_i^0 represents its reasoning embedding and c_i^0 its initial conclusion. This independence in initialization is crucial for establishing diverse starting points in the solution space.

Confidence Boundary Determination Following initialization, we adopt the bounded confidence mechanism from classical opinion dynamics models (Deffuant et al., 2000; Rainer and Krause, 2002). A confidence radius $R(t) = R_0 + \lambda \left(\frac{t}{T}\right)$ determines whether agents can consider opinions from each other, where R_0 is the initial radius and λ controls its temporal evolution. Two agents i and j can potentially interact only if their semantic distance falls within this radius: $E_{ij}^t = \mathbb{I}(d(s_i^t, s_j^t) \leq R(t))$, where $d(s_i^t, s_j^t)$ denotes the distance between agents' state and $\mathbb{I}(\cdot)$ is the indicator function. This bounded confidence mechanism helps prevent premature convergence while allowing the interaction scope to gradually expand as the debate progresses.

Weighted Opinion Exchange For agent pairs within confidence bounds, we compute influence weights based on both semantic similarity and answer conclusion consistency (See A.3). The overall influence weight incorporates this similarity measure along with agent-specific attributes:

$$w_{ij}^t = \beta_0 + \beta_1 \left(\frac{t}{T}\right) (1 + \gamma \sigma_i^t) \cdot \text{sim}(s_i^t, s_j^t), \quad (1)$$

where β_0 is the base confidence level, β_1 is the growth rate corresponding to debate progress, γ is the stability influence factor, σ_i^t denotes the agent's stability score and $\text{sim}(s_i^t, s_j^t)$ is the similarity score betweem agents' state.

These weights serve both topology and influence strength in regulating inter-agent interactions. Visibility of agent j's response to i is sampled according to the weight w_{ij}^t (w_{ji}^t if i to j), acting as the probability. Such adaptive directional topology

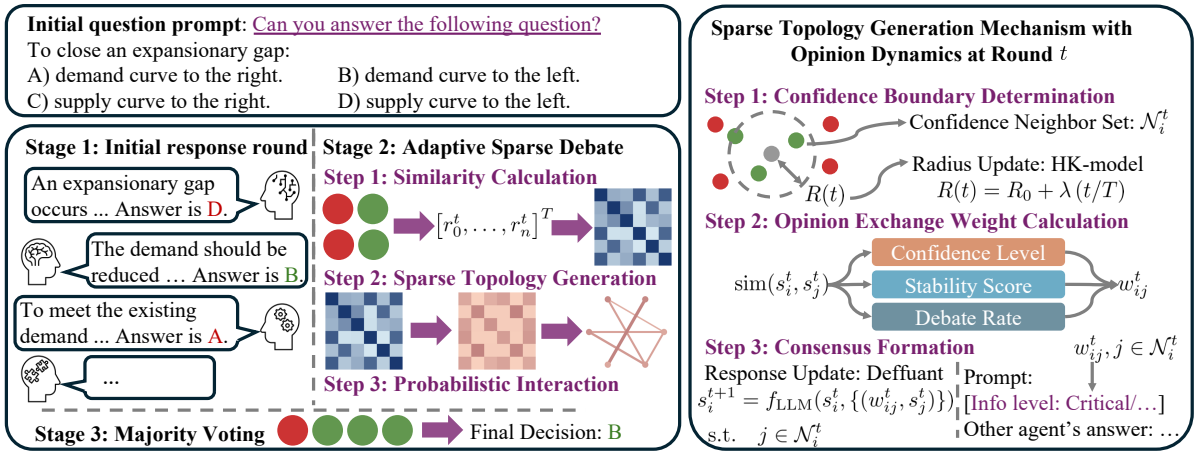

Figure 2: The process pipeline of ASMAD. Following S^2-MAD (Zeng et al., 2025), we adopts three stages in total. In the first stage, all agents gives the initial response. In the second stage, with proposed sparse topology generation mechanism, the agents are organized to debete with each other. In the last stage, the final decision is obtained via majority voting.

effectively reduces communication token overhead while preserving essential information flow paths.

Construction of agent prompts with varies with degrees of interaction strength, as practical workaround of opinion dynamics model in MAD scenarios. LLMs are prompted with one of: *Critical*, *Reference* and *Background* categories according to w if satisfied various thresholds (detailed in A.1).

Consensus Formation The consensus formation emerges through iterative debate rounds where agents continuously refine their positions through structured interactions:

$$s_i^{t+1} = f_{\mathrm{LLM}}(s_i^t, \{(w_{ij}^t, s_j^t) | j \in \mathcal{N}_i^t\}) \quad (2)$$

where $\mathcal{N}_i^t$ represents the set of visible agents to i at time t, and f_{LLM} denotes the language model's reasoning process. After sufficient rounds of debate, the final conclusion is determined through majority voting.

3.3 Framework Pipeline

The Adaptive Sparse Multi-Agent Debate (ASMAD) framework employs a structured three-stage pipeline as illustrated in Figure 2. Our methodology introduces dynamic opinion exchange mechanisms and adaptive topology generation to balance communication efficiency with debate effectiveness.

Stage 1: Independent Initialization The debate process begins with all agents independently generating initial responses to the given problem. Each agent i formulates its reasoning r_i^0 and conclusion c_i^0 without knowledge of other agents' perspectives, establishing diverse starting points across the solution space. As shown in the left panel of Figure 2, agents generate varied responses to questions like closing an expansionary gap, with conclusions spanning multiple possible answers.

Stage 2: Adaptive Sparse Debate The core of our approach lies in this intermediate stage, which orchestrates inter-agent interactions through three key steps:

1. *Similarity Calculation*: We compute semantic similarities between agent states using embedding distances and conclusion consistency, establishing a foundation for meaningful interactions.

2. *Sparse Topology Generation*: Based on the confidence boundary mechanism, we determine which agents can potentially interact. The confidence radius $R(t) = R_0 + \lambda\left(\frac{t}{T}\right)$ expands over time, gradually increasing the scope of potential interactions as the debate progresses.

3. *Probabilistic Interaction*: For eligible agent pairs, we calculate influence weights w_{ij}^t incorporating similarity measures, confidence levels, and stability scores. These weights determine both the probability of interaction and the influence strength when agents exchange opinions.

The right panel of Figure 2 details this mechanism, showing how confidence neighbors are determined, weights are calculated based on multiple factors, and how agents update their responses according to the Deffuant model. Importantly, the interaction prompt varies in intensity (Critical, Reference, or Background) based on the calculated weights, creating a natural gradient of influence in the form of text.

Stage 3: Consensus Formation In the final stage, after multiple rounds of adaptive debate, the framework aggregates individual conclusions through a majority voting mechanism. This democratic approach ensures that the final decision emerges from the collective wisdom of the agent ensemble rather than any single perspective. As depicted in the lower section of Figure 2, the voting process consolidates the diverse agent opinions into a single consensus answer (Final Decision: B).

4 Experiments

4.1 Tasks and Datasets

We mainly evaluate our framework on two benchmark datasets: GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2021),that either require multi-step reasoning or admit multiple valid solution paths while maintaining unambiguous answers. GSM8K presents grade school math word problems requiring step-by-step numerical reasoning. MMLU covers multiple-choice questions across various domains, where the challenge lies not only in answer format but in the diversity of valid reasoning approaches. We sampled 100 tasks from each dataset for agents to debate for 5 rounds as benchmark.

4.2 Model Configuration

To thoroughly evaluate the dynamic aspects and diversity benefits of our framework, we construct a heterogeneous agent population using three different LLM architectures: **LLaMA-3.1-8B-Instruct** (Touvron et al., 2023), **ChatGLM-4-9B-chat-abliterated** (GLM et al., 2024) and **Deepseek-math-7b-Instruct** (Guo et al., 2025). Each model type contributes 2 agents, resulting in a debate group of 6 participants. This configuration ensures sufficient diversity in reasoning approaches while maintaining manageable computational requirements. For practical deployment considerations, all deployed models leverage 4-bit blockwise quantization with mixed precision (Q4_K_M), enabling simultaneous execution of all 3 models on a single NVIDIA GeForce RTX 3090 GPU.

Task	Method	ACC	Token Cost (k/task)	Cost Saving
MMLU	**ASMAD(6,5)**	**73%**	**18.64**	**-64.2%**
	MAD(6,5)	49%	52.15	0
	S-MAD$_*$(6,5)	61%	26.42	-49.3%
	S-MAD$_o$(6,5)	54%	27.77	-46.8%
	GD(6,6)	53%	32.17	-38.3%
	S^{2}MAD(6,6)	46%	25.36	-51.4%
GSM8K	**ASMAD(6,5)**	80%	**21.94**	**-65.8%**
	MAD(6,5)	88%	64.08	0
	S-MAD$_*$(6,5)	83%	26.42	-58.8%
	S-MAD$_o$(6,5)	70%	27.77	-56.7%
	GD(6,6)	**90%**	32.17	-49.8%
	S^{2}MAD(6,6)	70%	25.36	-60.4%

Table 1: Performance of ASMAD and baselines across three tasks. Token cost is calculated as average of each topic debated. ASMAD significantly reduces token cost with comparable or improved accuracy. S-MAD with different structure is denoted as s(Star) and o(Ring).

4.3 Baseline and Evaluation Protocol

The primary baseline for comparison is MAD, a most straightforward fully-connected debate protocol without visibility control or prompt structuring. This baseline maintains complete information exchange between all agents throughout the debate process. We also take S-MAD, GD and S^2-MAD as comparable baselines with their best-claimed configurations.

Key evaluation metrics include: (1) Solution accuracy across different problem types; (2) Computational efficiency measured by token consumption. We further detailed ablation study and hyperparameters search of ASMAD in Appendix.4.5 and Appendix.C.

4.4 Main Results

Table 1 presents experimental results comparing ASMAD with the baseline MAD method. On MMLU, ASMAD demonstrates superior performance improvement while simultaneously reducing token consumption by 64.2%. Figure 3 and 4 shows ASMAD accelerates consensus with higher mean value and lower standard variance in similarity. On GSM8K, ASMAD delivers a remarkable 65.8% reduction in token cost with a slightly moderate accuracy drop. The performance gap may be attributed to similarity-based approach, which can be less effective than static topology methods when handling numerical answer discrepancies in mathematical reasoning tasks. It's worth noting that for GD(6,6) and S^{2}MAD(6,6) configurations,

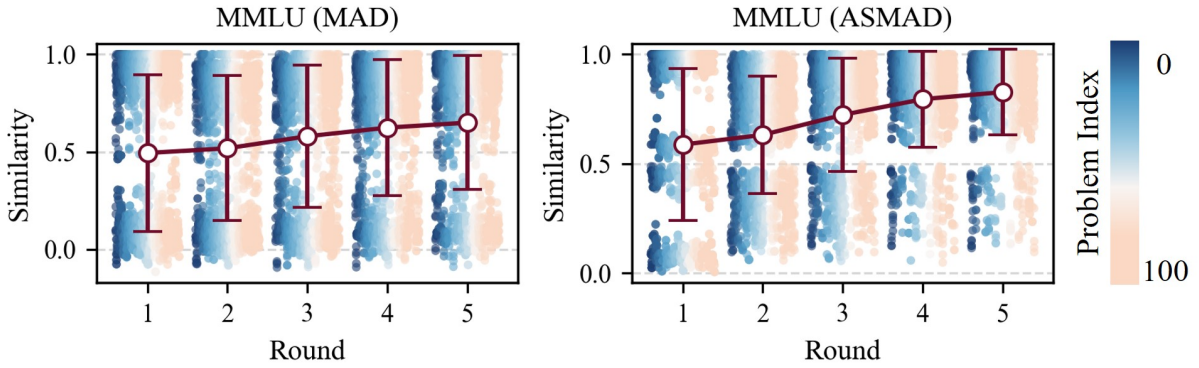

Figure 3: Similarity of agents vary toward consensus with increasing debate rounds where ASMAD provides better consensus rate (demonstrated in mean value and standard variance of similarity among agents) and speed

we implemented the best configuration claimed by the original authors (2 intra-group rounds + 1 inter-group round, repeated twice), which naturally introduces an additional round (6 vs 5) compared to other methods. This efficiency gain suggests that ASMAD mechanisms can effectively enhance reasoning capabilities, even among performance-limited quantized models.

To further assess the generalizability of ASMAD, we conducted additional experiments on the Graduate-level Physics Questions and Answers (GPQA) dataset and analyzed results in Appendix B.

4.5 Ablation study

Table 2 presents ablation study examining the contribution of each key module in ASMAD. The results clearly demonstrate that each component plays a crucial role in the overall performance of our approach with no single component being redundant. The complete ASMAD framework achieves the best accuracy-efficiency trade-off compared to any of its reduced variants. Results are collected through MMLU dataset.

Removing the trust radius mechanism leads to a significant drop in accuracy while only marginally reducing token costs. This confirms that trust radius effectively helps maintain a balance between exploration and exploitation during the adaptive debate process.

The balancing parameter λ, which calibrates the relative importance between semantic similarity and answer similarity, proves essential for ensuring comprehensive evaluation of agent contributions. Its removal results in a substantial accuracy degradation, highlighting the importance of considering both the reasoning process and the final answer when building consensus. It is more important in handling LLMs trained with different tune of response, even if 'temperature' parameters of them are configured identical.

Ablation Module	ACC	Token Cost (k/task)
ASMAD	**73%**	**18.64**
w/o trust radius	60%	17.95
w/o balancing parameter	62%	18.63
w/o weight clip	50%	12.62
w/o outlier filter	58%	14.83

Table 2: Ablation of key modules of ASMAD.

When weight clipping used in Appendix. A.2 is eliminated, we observe the most dramatic decline in accuracy, despite achieving the lowest token consumption. This suggests that while weight clipping may increase computational costs, it is fundamental for preventing premature convergence and maintaining solution quality.

The outlier filtering component also demonstrates its value, as its removal causes considerable accuracy reduction with only modest token savings. This confirms that identifying and mitigating the impact of extreme viewpoints contributes significantly to the robustness of the debate framework.

4.6 Performance with Larger Models

While our primary investigation focused on resource-constrained environments using quantized 7B-9B parameter models, we conducted additional experiments to evaluate whether ASMAD's ad-

Method	ACC	Token Cost (k/task)	Cost Saving
ASMAD(6,5)	**95%**	**37.16**	**-46.63%**
MAD(6,5)	90%	69.63	0.00%
S-MAD$_*$(6,5)	93%	45.71	-34.35%
S-MAD$_o$(6,5)	93%	46.96	-32.55%
GD(6,6)	93%	51.23	-26.43%
S^{2}MAD(6,6)	93%	46.96	-32.56%

Table 3: Accuracy and token cost comparison using larger models on MMLU

vantages persist when applied to larger language models. This exploration addresses an important question: does the adaptive sparse debate mechanism remain effective across model scales, or are its benefits limited to smaller, computationally restricted settings? For these experiments, we deployed a heterogeneous online ensemble consisting of three advanced models: DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025a), Doubao-1.5-thinking-pro (Seed et al., 2025), and DeepSeek-V3 (DeepSeek-AI et al., 2025b).

Table 3 presents the experimental results comparing ASMAD with baseline methods on the MMLU dataset. The findings are particularly noteworthy as they demonstrate that ASMAD's advantages become even more pronounced with larger models. ASMAD achieved the highest accuracy (95%) among all methods, outperforming the fully-connected MAD approach by 5 percentage points and other sparse debate methods by 2 percentage points.

Equally important, ASMAD maintained its substantial efficiency advantage with larger models, reducing token consumption by 46.63% compared to MAD. This reduction is especially significant considering that larger models typically generate longer outputs and incur higher computational costs per token. The token savings with ASMAD (37.16k tokens per task) compared to MAD (69.63k tokens per task) translate to substantial practical benefits in deployment scenarios.

When comparing ASMAD with other sparse debate methods, we observe a consistent pattern where ASMAD delivers superior accuracy while maintaining comparable or better efficiency. For instance, both star and ring topologies in S-MAD achieve 93% accuracy but consume more tokens than ASMAD (45.71k and 46.96k vs. 37.16k). Similarly, GD and S^{2}MAD reach 93% accuracy but with higher token costs (51.23k and 46.96k respectively).

4.7 Discussion and Key Findings

The experiments across different model scales and tasks provide comprehensive insights into the effectiveness and efficiency of our proposed ASMAD framework. This section synthesizes these findings to highlight the key advantages of ASMAD and its potential implications for multi-agent debate systems.

Balanced Performance Across Tasks Our results demonstrate that ASMAD achieves a remarkable balance between accuracy and computational efficiency across different task types. On MMLU, ASMAD delivered the highest accuracy (73%) among all methods while reducing token consumption by 64.2% compared to MAD. On GSM8K, while GD achieved the highest accuracy (90%), ASMAD (80%) maintained competitive performance with the largest token savings (65.8%). This pattern suggests that ASMAD's opinion dynamics approach is particularly effective for reasoning tasks with diverse solution paths, as in MMLU, while remaining competitive on more structured problems like GSM8K.

Scaling Benefits with Model Capability A particularly noteworthy finding is how ASMAD's advantages amplify when deployed with larger models. As shown in Table 3, ASMAD achieved 95% accuracy on MMLU with larger models while maintaining a substantial 46.63% token reduction compared to MAD. This indicates that as model capabilities increase, ASMAD becomes even more effective at leveraging their enhanced reasoning while controlling computational costs. The consistent pattern across both resource-constrained and larger model settings validates ASMAD's design principles as fundamentally sound.

Efficiency-Effectiveness Trade-off Throughout our experiments, we observed a consistent pattern where other sparse debate methods typically sacrifice either accuracy or efficiency compared to ASMAD. S-MAD topologies achieve moderate token savings but with lower accuracy, while methods like GD may match or exceed ASMAD's accuracy but with significantly higher token costs. ASMAD consistently delivers the most favorable trade-off, suggesting that its dynamic, adaptive approach to debate structure outperforms static topologies.

Component Importance The ablation studies in Table 2 reveal that each component of ASMAD

contributes significantly to its overall performance. The trust radius mechanism, balancing parameter, weight clipping, and outlier filtering all play crucial roles in maintaining ASMAD's accuracy. These findings highlight the importance of carefully balancing exploration and exploitation in multi-agent debates—allowing agents to consider diverse perspectives while preventing undue influence from outliers or premature convergence.

Practical Implications The significant token savings demonstrated by ASMAD (64.2% for MMLU and 65.8% for GSM8K with 7B-9B models; 46.63% with larger models) translate to substantial practical benefits. These include reduced computational costs, lower energy consumption, faster response times, and the ability to deploy effective multi-agent debate systems on more constrained hardware. Importantly, these benefits come with minimal or even positive impacts on accuracy, challenging the conventional wisdom that efficiency gains typically come at a performance cost.

In conclusion, ASMAD represents a substantial advancement in the design of multi-agent debate frameworks, offering a more principled approach to managing agent interactions through the lens of opinion dynamics. Its ability to maintain performance advantages across different tasks and model scales, combined with its significant efficiency improvements, positions ASMAD as a valuable approach for deploying multi-agent debate systems in a wide range of practical scenarios—from resource-constrained environments to high-performance computing settings.

5 Conclusion

This work introduces ASMAD, a novel framework that synergizes sociophysical opinion dynamics with MAD systems through two key innovations: (1) probabilistic semantic-guided attention gates that dynamically regulate opinion visibility based on textual reasoning similarity, and (2) a hybrid paradigm integrating adaptive trust-boundary regulation with opinion synchronization mechanisms. ASMAD enables efficient consensus formation through structured sparse interactions while preserving reasoning quality. Our experiments across multiple benchmarks demonstrate ASMAD's ability to significantly reduce token costs (by approximately 64-66% on MMLU and GSM8K) while maintaining or even improving accuracy compared to fully-connected MAD systems. The framework shows particularly strong performance with larger models, achieving 95% accuracy on MMLU with a 46.63% token reduction. These results establish that semantic-aware topology adaptation can simultaneously optimize reasoning quality and efficiency, making multi-agent debate more practical for real-world applications across different model scales.

Limitations

Our work, while demonstrating promising results, has several limitations worth acknowledging. While ASMAD demonstrates promising results, several avenues remain for future research. The primary limitation of our current approach is its sensitivity to hyperparameters, including confidence radius, growth rates, and similarity thresholds. The effectiveness of the adaptive debate protocol depends significantly on careful tuning of these parameters for specific tasks and agent configurations. To address this challenge, we plan to explore reinforcement learning approaches to dynamically tune these parameters based on debate context and agent behavior, potentially optimizing the diversity-consensus trade-off without manual intervention. Our initial validation with moderate-sized agent groups (6 agents) shows significant promise, though the dynamics and efficacy of our framework in larger debate clusters represents an intriguing direction for future investigation. The interplay between maintaining diverse perspectives and achieving efficient consensus at scale could yield valuable insights for multi-agent collaboration systems. Another promising direction is extending our semantic-guided approach to more diverse reasoning tasks, particularly those requiring specialized domain knowledge. While our current implementation shows strong performance on general reasoning tasks (MMLU) and mathematical problems (GSM8K), adapting the semantic similarity metrics for domain-specific applications could further enhance performance across specialized fields. We also acknowledge that adaptive consensus mechanisms could potentially amplify existing model biases or create information filtering effects. The selective information exchange mechanism, though efficient, requires careful implementation to avoid creating echo chambers where agents reinforce each other's misconceptions. Additionally, the framework's ability to generate more con-

vincing outputs through structured debate could be misused to produce more persuasive misinformation. Future work should explore techniques to detect and mitigate these effects while preserving ASMAD's efficiency advantages.

Ethical Considerations

In this research, Claude 3.5 Sonnet and Deepseek-R1 models are used as copilot, partially engaging in writing (sentence-level generations and grammar checking) and coding (fuzzing test and code-style polishing).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

DeepSeek-AI, Daya Guo, and Dejian et.al Yang. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, and Bei et.al Feng. 2025b. DeepSeek-V3 Technical Report. *Preprint*, arXiv:2412.19437.

Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. 2002. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4).

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3:87–98.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Seyed Rasoul Etesami and Tamer Başar. 2015. Game-theoretic analysis of the hegselmann-krause model for opinion dynamics in finite dimensions. *IEEE Transactions on Automatic Control*, 60(7):1886–1897.

Seyed Rasoul Etesami, Tamer Başar, Angelia Nedić, and Behrouz Touri. 2013. Termination time of multidimensional hegselmann-krause opinion dynamics. In *2013 American Control Conference*, pages 1255–1260. IEEE.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250.

Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.

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. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

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*.

Jan Lorenz. 2007. Continuous opinion dynamics under bounded confidence: A survey. *International Journal of Modern Physics C*, 18(12):1819–1838.

Luca Marconi and Federico Cecconi. 2020. Opinion dynamics and consensus formation in a deffuant model with extremists and moderates. *arXiv preprint arXiv:2010.01534*.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Hegselmann Rainer and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Artif. Societies Social Simul.*, 5:1–33.

ByteDance Seed, Yufeng Yuan, and Yu et.al Yue. 2025. Seed-Thinking-v1.5: Advancing Superb Reasoning Models with Reinforcement Learning. *Preprint*, arXiv:2504.13914.

Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280*.

Behrouz Touri and Cedric Langbort. 2014. On endogenous random consensus and averaging dynamics. *IEEE Transactions on Control of Network Systems*, 1(3):241–248.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Fatemeh Zarei, Yerali Gandica, and Luis Enrique Correa Rocha. 2023. Fast but multi-partisan: Bursts of communication increase opinion diversity in the temporal deffuant model. *arXiv preprint arXiv:2307.15614*.

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. *arXiv preprint arXiv:2502.04790*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2023. Exploring collaboration mechanisms for LLM agents: A social psychology view. *arXiv preprint arXiv:2310.02124*.

YunHong Zhang, QiPeng Liu, and SiYing Zhang. 2017. Opinion formation with time-varying bounded confidence. *PloS one*, 12(3):e0172982.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. 2022. Solving math word problems via cooperative reasoning induced language models. *arXiv preprint arXiv:2210.16257*.

A Implementation Details

A.1 Structured Information Exchange

The computed weights determine not only the influence strength but also how information is presented to each agent. We implement a three-tier prompt structure:

$$P_{ij}^t = \begin{cases} \text{[Critical]} & \text{if } w_{ij}^t > 0.40 \\ \text{[Reference]} & \text{if } w_{ij}^t > 0.25 \\ \text{[Background]} & \text{if } w_{ij}^t > 0.10 \end{cases} \quad (3)$$

This structured presentation helps agents prioritize information based on computed influence weights, while maintaining the natural language interaction paradigm of LLMs.

A.2 Self-confidence Evolution

The self-confidence of each agent evolves according to:

$$w_{ij}^t = \text{clip}(\beta_0 + \beta_1 \left(\frac{t}{T}\right)(1+\gamma\sigma_i^t) \cdot \text{sim}(s_i^t, s_j^t), w_{min}, w_{max})$$

where:

- β_0: base confidence level
- β_1: growth rate
- γ: stability influence factor
- σ_i^t: agent's stability score

A.3 Hybrid Similarity Computation

We introduce a novel similarity measure that combines reasoning process similarity and answer agreement:

$$\text{sim}(i,j) = \lambda \cdot \cos(r_i, r_j) + (1-\lambda) \cdot \mathbb{I}(c_i = c_j) \quad (4)$$

where:

- $\cos(r_i, r_j)$: cosine similarity between reasoning embeddings
- $\mathbb{I}(c_i = c_j)$: indicator function for answer agreement
- λ: balancing parameter

Method	ACC	Token Cost (k/task)	Cost Saving
ASMAD(6,5)	33%	30.38	-50.49%
MAD(6,5)	35%	61.37	0
S-MAD$_*$(6,5)	**39%**	35.36	-42.38%
S-MAD$_o$(6,5)	32%	36.00	-41.34%
GD(6,6)	35%	34.16	-44.34%
S^{2}MAD(6,6)	30%	**27.43**	**-55.31%**

Table 4: Experiment carried out on GPQA-Main dataset.

A.4 Stability Mechanism

The stability score for agent i at round t is:

$$\sigma_i^t = 1 - \frac{\sum_{k=2}^{t} \mathbb{I}_{c_i^k \neq c_i^{k-1}}}{t-1} \tag{5}$$

This score influences both self-confidence and inter-agent weights through the mechanisms described above.

A.5 Row Normalization

To ensure balanced influence distribution, we apply row normalization to the weight matrix:

$$\hat{w}_{ij}^t = \frac{w_{ij}^t}{\sum_k w_{ik}^t} \tag{6}$$

This normalized weight matrix $\hat{W}^t$ governs the information flow and influence dynamics in each round of debate.

A.6 Consensus formation

ASMAD enables agents to arrive at consensus faster. Figure 3 and Figure 4 show the dynamics of agent opinions through metrics of similarity.

B Extended Experiment on GPQA

To further assess the generalizability of ASMAD, we conducted additional experiments on the Graduate-level Physics Questions and Answers (GPQA) dataset, which poses significantly different challenges compared to MMLU and GSM8K. Table 4 presents the results of these experiments.

The results reveal an interesting deviation from the patterns observed with MMLU and GSM8K. While ASMAD maintained its substantial efficiency advantage (50.49% token reduction compared to MAD), it achieved slightly lower accuracy (33%) compared to MAD (35%) and S-MAD$_*$ (39%). This performance gap, though modest, warrants careful analysis to understand the domain-specific challenges of GPQA.

The GPQA dataset presents unique characteristics that likely influenced these results. Unlike MMLU and GSM8K, GPQA contains graduate-level physics questions that require specialized domain knowledge with strong interdependencies between reasoning steps. Physics problem-solving often involves multiple interconnected equations and principles that must be applied in a specific sequence, creating high dependency paths where early errors can significantly impact final answers. These problems also frequently involve complex symbolic manipulation and specialized notation that may challenge the semantic similarity metrics used in ASMAD for determining agent confidence and influence weights.

Several factors likely contributed to the observed performance pattern. ASMAD's confidence radius and influence weights depend on semantic similarity measures between agent reasoning. For highly specialized domains like physics, these measures may struggle to distinguish correct from incorrect reasoning paths when both use similar domain-specific terminology. The star topology of S-MAD$_*$, which performed best on GPQA, centralizes information flow through a hub agent, potentially advantageous when knowledge is unevenly distributed among agents. Additionally, the overall lower accuracy across all methods (30-39% compared to 70-95% on other datasets) suggests that GPQA presents inherent challenges for current LLM-based systems regardless of debate structure. In highly technical domains, early consensus formation based on similarity can sometimes reinforce plausible-sounding but ultimately incorrect approaches.

These findings do not indicate fundamental flaws in ASMAD's design but rather highlight opportunities for domain-specific adaptations. For highly specialized technical domains like physics, future improvements might include domain-specific similarity metrics that better capture the correctness of physics reasoning, adjustments to the confidence radius mechanism that account for the unique uncertainty characteristics of physics problems, and specialized agent roles based on demonstrated domain expertise. The GPQA results provide valuable insights into the boundary conditions of our approach, suggesting that while ASMAD excels at general reasoning tasks, highly specialized technical domains may benefit from domain-adapted variants.

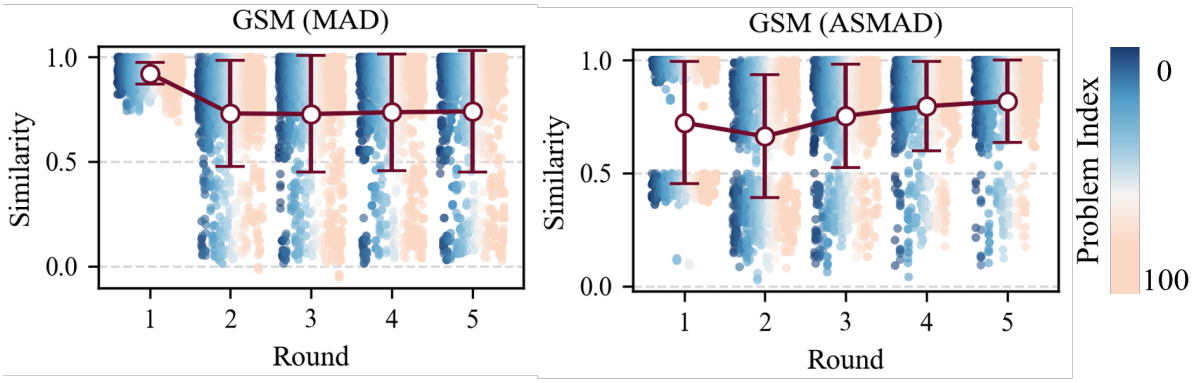

Figure 4: Similarity of agents vary toward consensus with increasing debate rounds where ASMAD provides better consensus rate (demonstrated in mean value and standard variance of similarity among agents) and speed

r_{init}\r_{final}	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**	**1.1**
0	63%	63%	63%	63%	63%	63%	63%
0.1	63%	63%	63%	67%	60%	60%	60%
0.2	70%	70%	**73%**	65%	60%	60%	60%
0.3	70%	70%	70%	65%	60%	60%	60%
0.4	70%	70%	70%	65%	60%	60%	60%
0.5	70%	70%	70%	65%	60%	60%	63%
0.6	70%	70%	70%	65%	60%	63%	60%
0.7	70%	70%	70%	65%	60%	63%	60%
0.8	70%	70%	70%	68%	60%	63%	60%
0.9	70%	70%	70%	67%	60%	60%	63%

Table 5: Searching init and final configurations of trust-boundary of ASMAD. Trust-boundary dynamically changes from r_{init} to r_{final} according to debate progress.

λ\ACC	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**
0	69%	69%	65%	63%	62%
0.1	70%	71%	68%	67%	66%
0.2	70%	68%	70%	70%	**72%**
0.3	70%	68%	68%	68%	69%
0.4	66%	64%	65%	62%	62%
0.5	67%	65%	66%	66%	66%
0.6	67%	65%	65%	65%	65%
0.7	69%	68%	68%	67%	68%
0.8	69%	68%	67%	66%	67%
0.9	69%	65%	65%	64%	64%
1	69%	65%	65%	64%	64%

Table 6: Searching balancing parameter λ of ASMAD.

C Hyper-parameter Search

As shown in Table 5, Table 7 and Table 6, we performed linear search for hyper-parameters defined in Section 3.2 of ASMAD.

D Prompt

As shown in Table 8 and Table 9, we use identical prompt configuration with baseline methods (Zeng et al., 2025; Liu et al., 2024) to preserve fairness in experiment.

w_{min}	w_{max}	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**
0	**0.1**	67%	0%	33%	33%	20%
	0.2	67%	37%	53%	57%	47%
	0.3	67%	37%	53%	50%	43%
	0.4	67%	40%	57%	50%	37%
	0.5	67%	43%	57%	53%	50%
	0.6	67%	43%	57%	57%	57%
	0.7	67%	43%	57%	63%	53%
	0.8	67%	43%	53%	60%	50%
	0.9	67%	47%	57%	63%	50%
0.1	**0.2**	67%	63%	63%	60%	60%
	0.3	67%	63%	63%	60%	60%
	0.4	67%	63%	67%	60%	57%
	0.5	67%	63%	63%	60%	60%
	0.6	67%	63%	63%	60%	60%
	0.7	67%	63%	63%	60%	60%
	0.8	67%	63%	63%	60%	63%
	0.9	67%	63%	63%	57%	57%
0.2	**0.3**	67%	60%	53%	57%	53%
	0.4	67%	60%	57%	53%	57%
	0.5	67%	60%	57%	57%	60%
	0.6	67%	60%	57%	57%	57%
	0.7	67%	60%	60%	60%	60%
	0.8	67%	60%	60%	60%	60%
	0.9	67%	60%	60%	60%	60%
0.3	**0.4**	67%	60%	63%	63%	63%
	0.5	67%	60%	67%	67%	**70%**
	0.6	67%	60%	**70%**	**70%**	**70%**
	0.7	67%	60%	63%	63%	60%
	0.8	67%	60%	60%	63%	63%
	0.9	67%	60%	60%	63%	67%
0.4	**0.5**	67%	67%	67%	67%	63%
	0.6	67%	67%	67%	67%	67%

Table 7: Searching range of weight clip of ASMAD. This search ends at $w_{min} = 0.4$ to ensure $w_{min} \leq w_{max}$ as reasonable clipping.

Table 8: Prompts in Each Stage. List of prompts used in each task.

Type	Task	Prompt
System	All	Welcome to the debate! You are a seasoned debater with expertise in succinctly and persuasively expressing your viewpoints. You will be assigned to debate groups, where you will engage in discussions with fellow participants. The outcomes of each group's deliberations will be shared among all members. It is crucial for you to leverage this information effectively in order to critically analyze the question at hand and ultimately arrive at the correct answer. Best of luck!
Starting	GSM8K	Can you solve the following math problem? <Problem> Explain your reasoning. < Output Format >.
	MMLU	Can you answer the following question? <Problem>: A) , B) , C) , D) Explain your answer. <Output Format>.
	GPQA	Can you answer the following question? <Problem>: A) , B) , C) , D) Explain your answer. <Output Format>.
Intra-group Debate	All	These are the recent unique opinions from other agents that differ with yours: <other agent responses> Using the opinions carefully as additional advice, can you provide an updated answer? Examine your solution and that other agents step by step. <Output Format> .
Summary	All	These are the recent/updated and unique opinions from all agents: <all agent responses> Summarize these opinions carefully and completly in no more than 80 words. Aggregate and put your final answers in parentheses at the end of your response.
Inter-group Debate	All	These are the recent unique opinions from all groups: one group responses: <group summary>. Using the reasoning from all groups as additional advice, can you give an updated answer? Examine your solution and that all groups step by step. <Output Format>.

Table 9: Output Format Requirements in Each Dataset.

Task	Output Format Requirements
GSM8K	Your final answer should be a single numerical number, in the Form \boxed{{answer}}, at the end of your response.
MMLU	Put your final choice in parentheses at the end of your response.
GPQA	Put your final answer in the Form \ The correct answer is (insert answer here).