Building Efficient Heterogeneous Multi-Agent Debate Frameworks for Factual LLMs

Ethan Justice[1], **Satyak Khare**[1], **Wenjia Lu**[1], **Daniel Vega**[1], **Eli Wiegman**[1], **Christopher Zhou**[1]

[1]University of Michigan
{ethanjus, satyakkh, wenjialu, danivega, eliw, zhoucz}@umich.edu
Group 13

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. However, they remain prone to logical inconsistencies and hallucinations, particularly when challenged with complex multi step reasoning tasks. To address this, the Multi Agent Debate (MAD) framework emerged, where multiple model instances engage in conversational debate to refine responses. This approach serves as an inference time optimization strategy that requires no architectural modifications or weight updates. In this work, we first replicate the findings of the foundational homogeneous Multi Agent Debate framework (Du et al. 2024) using Meta-Llama-3.1-8B. Our replication reveals that this framework provides performance improvements on MMLU (18%) and GSM8K (Grade School Math) (20%), but has a negligible performance decrease on the Biography benchmark (-1%).

The performance gain comes at an extreme computational cost that increases exponentially with the number of agents and rounds. Also, homogeneous agents tend to form "echo chambers" where identical model instances amplify shared biases rather than correcting them (Ye et al. 2025). To mitigate these limitations, we propose a novel extension that combines confidence based early stopping with heterogeneity. Our architecture integrates the "Debate Only When Necessary" (DOWN) framework with heterogeneous debating agents. Unlike traditional setups that trigger debate for every query, our system utilizes a generalist model for initial screening. The computationally expensive multi agent debate triggers only when model confidence falls below a calibrated threshold. This approach optimizes the trade off between accuracy and efficiency, ensuring resources are allocated only to complex queries while guaranteeing diverse model perspectives. We conclude that while heterogeneity successfully mitigates hallucination reinforcement in knowledge tasks, relying on internal logits for logic gating creates a "syntactic blind spot" that necessitates external verifiers for robust mathematical reasoning. These findings suggest that future efficient multi agent systems must decouple reasoning verification from token generation confidence.

Introduction

Problem Significance

A fundamental limitation of Large Language Models (LLMs) is their reliance on a reasoning path token-by-token and rarely self-correct during multi-step logical reasoning. This issue is compounded by the probabilistic nature of their architecture. When prompted, models frequently engage in "post-hoc rationalization," generating plausible-sounding justifications that may not align with their internal decision making processes. This greatly harms domain-specific use where factuality and accuracy are paramount, in areas such as education, healthcare, and public policy.

The authors (Du et al. 2024) addresses this lack of reliability in LLMs by reframing LLM improvement as a problem of collective deliberation rather than individual cognition. Their idea that multiple independent LLMs can reason, debate, and converge toward better answers offers a fresh direction for improving factuality and reasoning without retraining or architectural adjustments.

This mirrors human collective reasoning, where group deliberation corrects individual biases and enhances truth-finding. The significance here is conceptual: it proposes that reasoning and truthfulness in AI may emerge not from deeper neural capacity, but from structured social interaction among equally capable models. This approach aligns with a broader movement toward transparent, multi-perspective AI, which may become foundational for building systems that a society can trust to reason, not merely to predict.

Key Research Problem

Inspired by the rapid growth of multi-agent systems (MAS) due to their flexibility and intelligence (Balaji and Srinivasan 2010)(Dong 2024), this paper addresses the problem of how to design efficient and robust multi-agent reasoning systems for modern LLMs. While prior work on MAD shows substantial improvements in factuality and reasoning, these gains come with large computational overhead and rely on homogeneous agents that share the same limitations (Ye et al. 2025). At the same time, recent work on conditional debate (Eo et al. 2025) suggests that much of this debate is unnecessary, and work on heterogeneous multi-agent systems (Ye et al. 2025; Zhang et al. 2025) demonstrates that diversity across agents is crucial for state of the art performance. However, no existing framework integrates these insights. We therefore investigate two research questions:

1. **Replication**: Does the original MAD framework still yield meaningful improvements in factual accuracy and reasoning capabilities on contemporary LLMs?

2. **Extension**: Can we design a new MAD system, combining confidence-gated debate with heterogeneous agents, that achieves superior factuality and reasoning at lower computational cost?

Extension Overview

While the MAD framework may be effective, it is not without its flaws of being computationally expensive and having room for performance gains. Our replication data confirms this substantial overhead, revealing that the MAD protocol increased computational costs by approximately 11 to 15 times (in total FLOPs) compared to the single-agent baseline across all benchmarks. Additionally, homogeneous agents suffer from correlated errors and echo chambers (Ye et al. 2025)(Zhang et al. 2025). We hypothesize that (1), debate is not necessary for every query, allowing us to reduce computational costs associated with debate rounds and (2), that using a more diverse set of debate agents can allow agents to cover each other's blind spots and lead to even more solid reasoning and factuality. Our extension therefore proposes a system where we use a confidence threshold to trigger debate only when necessary and employ heterogeneous agents to facilitate such debate, aiming to address the core limitations of the original MAD paper.

Related Work

Multi-Agent Debate and Consensus. The field of reliable LLM reasoning has shifted from single-agent prompting to Multi-Agent Systems (MAS), where multiple model instances collaborate or debate to refine answers. Du et al. (Du et al. 2024) established the foundational framework, demonstrating that multi-agent consensus can significantly reduce hallucinations compared to single-agent generation. However, recent meta-analyses have challenged the universality of the "more agents is better" assumption. Zhang et al. (Zhang et al. 2025) conducted a systematic evaluation of five representative MAD methods across nine benchmarks, revealing that homogeneous debate often fails to outperform simple single-agent baselines (like Self-Consistency), suffers from "echo chamber" effects where identical models reinforce shared errors, and consumes significantly more computational resources. In contrast, our work acknowledges these limitations by using the Du et al. framework only as a baseline, and directly addresses the issues identified by Zhang et al. (Zhang et al. 2025) through architectural changes within our proposed extension.

Heterogeneity in Multi-Agent Systems. Building on the need for diversity, other researchers (Ye et al. 2025) introduced the X-MAS paradigm, a framework for "Heterogeneous LLM-driven MAS". Unlike traditional frameworks that rely on a single model to drive all agents, X-MAS assigns specialized models to specific roles (e.g., Planner, Reasoner, Reviewer) based on empirical benchmarking. Their extensive empirical study (X-MAS-Bench) demonstrated that replacing homogeneous agents with diverse, specialized models could yield up to a 47% performance boost on complex reasoning datasets like AIME, without requiring structural redesigns of the interaction graph. Our paper explicitly adopts this hypothesis: unlike our baseline replication using identical Llama-3 agents, our extension leverages the X-MAS findings to demonstrate that employing diverse agents yields stronger reasoning than our baseline replication of homogeneous agents.

Efficiency and Adaptive Execution. A major barrier to the adoption of debate frameworks is the significant computational cost. Addressing this, Eo et al. (Eo et al. 2025) proposed "Debate Only When Necessary" (DOWN), an adaptive framework that utilizes a confidence-based gating mechanism: if an agent's initial response exceeds a confidence threshold, the system bypasses the debate phase entirely. They demonstrated that this selective activation can reduce computational overhead by up to six times while maintaining accuracy. This confirms that the debate process, while powerful, is often redundant for simpler queries. Our work bridges this efficiency with the accuracy benefits of heterogeneity. By integrating DOWN with a heterogeneous agent structure, we aim to create a system that optimizes the trade-off between computational efficiency and factual accuracy.

Source Code:
https://github.com/Wenjia-Lu/llm_multi_agent

Background

Multi-Agent Debate and Consensus

The MAD framework can be demonstrated in the figure below.

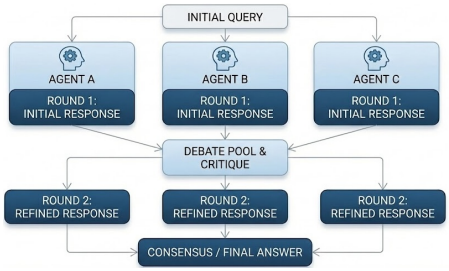

Figure 1: Visual Diagram of the Multi-Agent Pipeline

The Efficiency-Accuracy Trade-off

While MAD has shown to improve accuracy, increasing arithmetic performance from 67% to 82% in initial studies, it introduces a significant computational bottleneck. Recent analyses indicate that while scaling the number of agents and debate rounds can push accuracy near 98%, it can result in a token cost increase of over 101x (Zhang et al. 2025). This diminishing return on investment highlights the need for frameworks that debate selectively rather than exhaustively.

Advanced Debate Architectures

To mitigate the efficiency overhead while retaining accuracy gains, we incorporate mechanisms from two advanced frameworks in our extension:

1. **Dynamic Confidence-Guided Debate (DOWN):** Proposed to reduce redundancy, this method introduces a confidence threshold. A relatively lightweight agent generates an initial response and a confidence score (derived

from token log-probabilities). If the confidence exceeds a certain threshold, the system terminates immediately, bypassing the debate phase. Debate is only triggered for low-confidence queries, effectively filtering out "easy" problems from the expensive multi-agent pipeline. (Eo et al. 2025)

2. **Heterogenous Multi Agent Debate:** Recent evaluation shows that traditional debate frameworks underperform because all agents are identical copies of one model, causing correlated errors. The proposed heterogenous MAD architecture introduces model diversity by sampling agent responses from multiple foundation models. This yields consistent accuracy gains across benchmarks by allowing agents with different strengths to correct one another, especially in cases where one model fails and another succeeds. (Ye et al. 2025)

Models

A primary challenge in replicating the findings of the original paper was the deprecation of LLM used, GPT-3.5-turbo-0301. Thus, we attempted to use GPT-3.5-turbo, which did not yield strong enough results to use as a baseline for our paper. We decided to prioritize the selection of open-source models to rigorously test the debate framework's capacity to enhance the reasoning of accessible, resource-efficient architectures. By utilizing Meta-Llama-3.1-8B, we aim to demonstrate that the MAD mechanism can serve as a powerful inference-time optimization, effectively allowing smaller, locally hosted models to overcome their parameter limitations. Unlike proprietary flagship models, which function as "black boxes," open-weights models provide a transparent and fully reproducible testing ground, ensuring that any performance gains are attributable to the debate framework rather than hidden API updates.

To validate the framework's generalizability, we established a hybrid testing environment. The core of our replication focuses on the Meta-Llama-3.1-8B model to evaluate the hypothesis that debate can elevate the performance of local agents to rival larger systems without requiring massive computational resources. This allowed us to benchmark the progress of modern open-source agents against the standard established in the original literature.

Datasets

We utilize three primary benchmarks to evaluate reasoning and factuality, mirroring the original MAD paper:

1. **MMLU (Massive Multitask Language Understanding):** Evaluates general world knowledge and problem-solving ability across 57 subjects with multiple choice questions.
2. **GSM8K (Grade School Math):** A benchmark of high-quality linguistically diverse grade school math word problems, specifically chosen to test multi-step mathematical reasoning.
3. **Biographies Benchmark:** A specialized dataset utilized in Du et al. (2024) consisting of biographical data on people to specifically measure hallucination rates in generative text. The data was taken from Wikipedia.

Performance Metrics

We evaluate our hybrid pipeline using two dimensions:

1. **Accuracy:** The percentage of correct responses across benchmarks.
2. **Computational Efficiency:** To control for model size differences, we measure the cost of inference. We estimate Floating Point Operations (FLOPs) using the standard approximation $C \approx 2N \cdot T_{total}$, where N is the parameter count, and T_{total} (Kaplan et al., 2020). This allows us to compare compute used to generate each answer, controlling for model size and total token usage.

Methodology, Results, and Discussion

Our replication framework relies on the core logic provided by the original MAD paper (Du et al. 2024). To adapt its cloud-oriented design for high-performance computing (HPC) environment, we created a custom inference wrapper that intercepts agent generation calls and routes them to a locally hosted vLLM instance on the Great Lakes cluster. Thus, we preserved the original debate structure, prompts, and consensus mechanism while running the open-source Meta-Llama-3.1-8B model. We used the 4-bit AWQ-quantized variant to fit within GPU memory limits, and the wrapper handled context-window management (4096 tokens) and memory buffers to ensure stable performance.

The original framework was evaluated against three baseline methods: Single Agent, Single Agent with Reflection, and Multi-agent with Majority. These methods were tested across a mix of generated datasets (e.g., Arithmetic) and established benchmarks (e.g., MMLU and GSM8K). The results were significant: compared to the baseline model of using a single agent, multi-agent debate enhances arithmetic accuracy by 15% (from 67% to 82%), while performance on MMLU and GSM8K (Grade School Math) benchmarks increased by 7% respectively.

However, during our replication process, we encountered obstacles rooted in the deprecated model (GPT-3.5 turbo-0301) originally used in the MAD paper. Our replication process opted for GPT-3.5-turbo and Meta-Llama-3.1-8B instead. Unlike the static version originally used, the GPT-3.5-turbo API represents a continuously updated model heavily optimized for user alignment. Our initial results with the GPT-3.5-turbo model did not yield strong enough results to use as a baseline for our paper, likely due to 'model drift', where model updates and increased safety fine-tuning change the models behavior, suppressing contentious interactions required for debate (Table 5 in Appendix). This motivated our shift to a static, open-source setup, which removes the opacity of proprietary API changes and ensures our metrics reflect the debate framework itself rather than external model fluctuations.

Results Comparison

By using an open-weights model, we gain full control over the inference parameters. We found the existing framework produced more persuasive results aligned with the declared standards set by the paper.

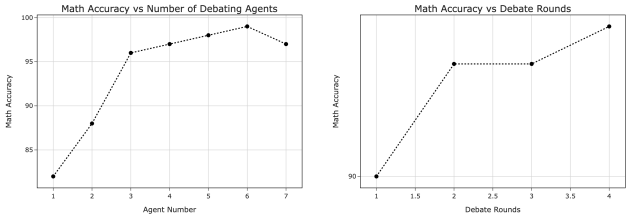

Figure 2: Replication Number of Debating Agents & Debate Rounds

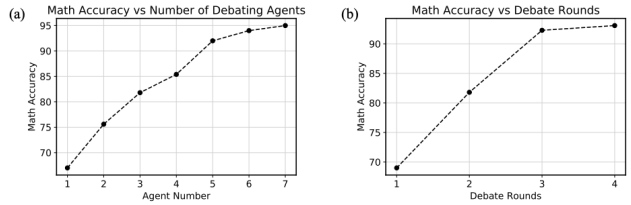

Figure 3: MAD Research Paper Number of Debating Agents & Debate Rounds

Figures 2 and 3 illustrate key performance metrics from our replication, providing a comparison to the original MAD study. Although the results demonstrate similarities, differences enforce model inconsistencies. Our replication results have a significantly higher floor. The jump from two debating agents to three is drastic for replication, whereas its MAD counterpart is subtle, as the debate rounds have a range of 4.0 instead of 24. However, the results fail to recognize an underlying setback: Homogeneous Multi-Agent Debate. The framework has the limitation that agents are instances of one underlying model. Use of a homogeneous framework produces a limited ceiling, as there is insufficient randomness. This conclusion is solidified by side-by-side comparisons in MMLU, GSM8K, and Biography benchmarks.

Table 1: Comparison of accuracy scores between the Single-Agent baseline (1/1) and the Multi-Agent setup (3 agents, 2 rounds) across three benchmarks.

Agents/Rounds	MMLU	GSM8K	Bio
1/1	48%	76%	55%
3/2	68%	80%	54%

Among the three categories, the multidisciplinary multiple-choice dataset is the only one that demonstrates significant accuracy improvement: +20% gain from 48% to 68% when transitioning from single-agent to multi-agent. Since MMLU has a large subject variety (57 distinct subjects), A single agent might have a "blind spot" or miss a nuance in a niche subject question, whereas, in a multi-agent debate, different agent instances effectively pool their latent knowledge. In contrast, Grade School Math is a closed system of logic, improving only 4%, from 76% to 80%. If the model knows how to perform the algorithm, the individual numbers matter less, and there's a reduced chance of a blind spot. There are fewer obscure facts to miss, so pooling knowledge yields diminishing returns.

Surprisingly, on the biography dataset where models are asked to generate a list of dates and notable events, MAD actually reduced performance by 1%, from 55% to 54%. A plausible explanation for the Biography results lies in the recursive nature of the evaluation, where the model is tasked with grading its own outputs. Furthermore, Meta-Llama-3.1-8B, with 8 billion parameters likely suffers from an inability to filter stochastic noise effectively, struggling to differentiate between true qualitative improvements and mere phrasing changes during the scoring process.

Table 2: Comparison of computational cost (FLOPs) between Single-Agent and Multi-Agent frameworks across three benchmarks.

Benchmark	Single	Multi
Biography	0.13P	2.01P
MMLU	0.16P	1.70P
GSM8K	0.14P	1.97P

To rigorously quantify the computational cost inherent to the debate framework, we tracked the total Floating Point Operations (FLOPs) accumulated across all inference passes. Our computational analysis revealed that the Multi-Agent Debate protocol introduces a massive overhead compared to the Single-Agent baseline, increasing compute usage by an order of magnitude (from 10^{14} to 10^{15} scale) across all benchmarks. For instance, the Biography task saw FLOPs surge from 1.30×10^{14} in the single-agent baseline to 2.01×10^{15} in the multi-agent setup, nearly a 15-fold increase in computational cost. Similar scaling factors were observed in MMLU (1.58×10^{14} to 1.70×10^{15}) and GSM8K (1.44×10^{14} to 1.97×10^{15}), confirming that while debate may refine reasoning, it does so at a prohibitive resource cost that necessitates the efficiency-focused extensions proposed.

Interpretation of Disagreement: While our replication confirmed the utility of debate for logical reasoning (GSM8K), our results for the Biography task (a -1% degradation) starkly contradict the findings of Du et al. (2024), who reported reduced hallucinations. We hypothesize that this discrepancy is driven by the "Echo Chamber" effect inherent to smaller, homogeneous models. Unlike the larger proprietary models used in the original study, the 8B-parameter agents lack the discriminatory power to distinguish between a peer's "plausible hallucination" and a "factual correction." Consequently, instead of filtering out errors, the consensus mechanism reinforces them, leading to hallucination propagation rather than mitigation.

Computational Viability: Our analysis of the computational overhead reveals a critical scalability issue. As detailed in the results, the multi-agent protocol required 2.01×10^{15} FLOPs for the Biography task, compared to just 1.30×10^{14} for the single-agent baseline. This order-of-magnitude

increase in resource consumption yields diminishing (or negative) returns for simpler queries, suggesting that the "always-on" debate structure proposed by the original paper is computationally inefficient for general-purpose deployment.

Extension

The original MAD framework demonstrates that having multiple identical language-model agents argue and converge on an answer can improve factual accuracy and multi-step reasoning. However, these gains come with substantial additional compute costs caused by multiple agents and rounds. Moreover, we observed that model deliberation sometimes leads to stagnant or even degraded performance compared to single-shot responses on simpler queries, where the consensus mechanism may enforce a common hallucination rather than correct reasoning.

These observations motivate the design of a more selective and diverse multi-agent reasoning pipeline. We introduce an extension that integrates two advanced multi-agent system (MAS) concepts: confidence-gated collaboration and heterogeneous agent roles. This approach is designed to preserve the accuracy advantages of debate while mitigating the computational overhead that hinders classical MAD, effectively enforcing a "Debate Only When Necessary" (DOWN) protocol (Eo et al. 2025).

Methodology

Our extended framework operates in a four-step pipeline: (1) Initial Response & Confidence Assessment, (2) Debate Engagement Check, (3) Heterogeneous Debate (conditional), and (4) Final Aggregation.

Step 1: Initial Response and Confidence Scoring We employ Llama-3.1-8B as the "Gate" agent. For a given query q, the Gate generates an initial response, r_1.

Simultaneously, we compute a scalar confidence score c_1 to quantify the model's internal certainty regarding its answer. Following the method described in recent literature (Eo et al. 2025), we define c_1 as the arithmetic mean of the token probabilities over the generated sequence. For a response consisting of tokens $t_1, \ldots, t_N$, the confidence is calculated as:

$$c_1 = \frac{1}{N} \sum_{i=1}^{N} P(t_i) \tag{1}$$

where $P(t_i)$ is the softmax probability of the token generated at step i. This metric captures the model's average certainty across the entire reasoning trace and final answer.

Step 2: Debate Engagement Check The system compares the calculated confidence c_1 against a predefined threshold θ. Based on empirical tuning on a validation set, we established $\theta = 0.8$, a hyperparameter adopted directly from the initial DOWN framework findings (Eo et al. 2025) which balances false positives and computation savings. The system deems the answer reliable if the confidence is greater than the threshold,

$$c_1 > \theta \ (0.8) \tag{2}$$

Step 3: Heterogeneous Debate Protocol When debate is triggered, we diverge from the homogeneous architecture of standard MAD. Instead of identical agents, we initialize a panel of three heterogeneous models chosen with orthogonal specialties to maximize perspective diversity based of of benchmark performance (Eo et al. 2025) (Zhang et al. 2025). Llama-3.1-8B was chosesn as a generalist agent (73.0% performance on MMLU) DeepSeek-R1-Qwen-7B was chosen for its semantic reasoning (49.1% performance on GPQA-Diamond) and Mathstral-7B: was chosen as a math expert (77.1% on GSM8K) (Mistral AI Team 2024)(DeepSeek-AI 2025)(Llama Team 2024)

We utilize a Parallel Generation with Cross-Seeding protocol. In the first round, all three agents generate responses to the query independently. In the second round, the context window for each agent is updated to include the responses generated by the other two agents in the previous round. The agents then regenerate their answers, allowing them to critique or adopt the reasoning of their peers. This process runs for a maximum of 2 rounds.

Step 4: Aggregation and Metrics Final answers are extracted using structured parsing of boxed expressions. The final prediction is selected via majority voting among the three agents. We evaluate factual accuracy against ground truth using GPT-4o as an evaluator and log computational metrics including token count and FLOPs.

Results and Analysis

We evaluated the extension on the Biography dataset (knowledge retrieval) and the GSM8K dataset (mathematical reasoning). The MMLU benchmark was excluded as its core functionally overlaps with Biography in testing knowledge retrieval, and we prioritized analyzing open-ended generation over multiple-choice classification. The results highlight a distinct trade-off between computational efficiency and calibration reliability across different domains.

Efficiency Gains in Knowledge Retrieval (Biography) On the Biography dataset, the confidence gate functioned as intended, effectively filtering "easy" queries from "hard" ones. Out of 25 processed queries, the Gate accepted 11 (44%) as reliable and sent 14 (56%) to debate.

Table 3: Compute Usage Summary (Biography)

Category	Count	Avg Tokens	Avg FLOPs	FLOPs
Gated	11	386.9	6.19 T	68.10 T
Debated	14	5,043.7	75.31 T	1.05 P
Overall	25	2,994.7	44.90 T	1.12 P

As shown in Table 1, debated questions consumed approximately 12.17x more compute than gated questions. By filtering 44% of the traffic, the gating mechanism achieved a 40.4% reduction in total compute compared to a baseline where all questions are debated. Importantly, this efficiency did not compromise performance. The accuracy for Gated responses (0.7302 ± 0.056) was statistically equivalent to the Debated responses (0.7308 ± 0.087), resulting in a combined accuracy of 0.7303. It is notable that the Gated responses achieved 73% accuracy compared to the single-agent baseline of 55% reported in our replication, while using an iden-

tical model. This improvement is not due to the model becoming smarter, but due to selection bias: the confidence gate successfully identified and routed the "easy" queries to the single agent, while filtering out hard queries that typically lower the average score. This confirms that for general knowledge tasks, the model's internal confidence is a reliable proxy for correctness.

Mathematical Reasoning Failure (GSM8K): In strong contrast to the Biography results, the system failed to trigger debate on the GSM8K math dataset. In our sample of 10 GSM8K questions, 100% of the initial responses scored a confidence $c_1 > 0.9$, well above the 0.8 threshold. Consequently, the system bypassed debate for every question, resulting in a final accuracy of only 30.0% (3/10). This phenomenon exposes a critical flaw in using raw token probabilities as a proxy for semantic confidence in mathematical domains. We hypothesize that this is due to "Syntactic Determinism." In mathematical generation, the model often assigns high probability to tokens based on the rigid formatting of equations (e.g., operator placement, currency symbols, comma separators) rather than the correctness of the underlying calculation. We isolated a compelling example of this behavior in Question #3, where the model attempts to calculate the post-repair value of a house ($80,000+50,000$). The correct sum is $130,000$, but the model outputs $120,000$. Despite this arithmetic error, the model assigns a probability of 1.0 to the incorrect digits, as shown in the selected token trace shown in Table 4.

Table 4: Token Probability Trace for Incorrect Arithmetic (Question #3). Context: "This is $80,000 + $50,000 =..."

Token	Probability	Note
+	1.000	Syntactically determined
$	1.000	Syntactically determined
50	1.000	Operand
,	1.000	Formatting
000	1.000	Formatting
=	1.000	Operator
$	1.000	Currency Symbol
120	**1.000**	**Arithmetic Error** ($80 + 50 \neq 120$)
,	1.000	Formatting
000	1.000	Formatting

As illustrated, the confidence is saturated with 1.0, with a final confidence score for the answer being 0.982384, even though the answer is incorrect.

Throughout the entire erroneous sequence, the model appears to be "confident" in the format of the equation, predicting that a number must follow the equals sign, and that commas must follow thousands, masking the fact that the semantic reasoning is flawed. Because the confidence score, c_1, is an average of these probabilities, the high certainty of the syntactic tokens (",", "$", "=") inflates the overall score, rendering the gate ineffective for detecting reasoning errors. Crucially, this failure was architectural, not pedagogical. The specialized Mathstral agent never had the opportunity to correct the error because the "Gatekeeper" (Llama-3.1-8B-Instruct) confidently hallucinated the syntax, preventing the debate from triggering.

Conclusions

Our replication of the MAD framework highlights that while multi-agent debate can improve reasoning, its efficacy is highly context-dependent. We observed that general knowledge tasks benefited most from agent deliberation, as agents effectively pooled complementary knowledge to correct blind spots. Conversely, closed-system logical tasks showed diminishing returns, and generative fact-recall occasionally suffered from degradation. Across all benchmarks, the homogeneous debate architecture introduced substantial computational overhead and suffered from a performance ceiling caused by correlated failure modes among identical agents.

To address these inefficiencies, our extension integrated confidence-gated collaboration and heterogeneous agent roles, utilizing Llama-3.1 alongside specialist models like DeepSeek-R1 and Mathstral. This hybrid approach successfully maintained accuracy while reducing computational costs by over 40% in knowledge-heavy domains. By enforcing a "Debate Only When Necessary" protocol, the system demonstrated a significantly improved accuracy-per-token ratio compared to the original framework.

However, a critical limitation emerged during mathematical evaluation. Our analysis of the GSM8K dataset revealed that current LLMs exhibit "Syntactic Determinism," where models assign inflated confidence to incorrect outputs simply because they adhere to valid mathematical formatting. As demonstrated in our results, a model can hallucinate arithmetic errors (e.g., $80,000 + 50,000 = 120,000$) with 100% token probability. This finding suggests that while confidence gating is highly effective for knowledge retrieval, future reasoning systems must look beyond simple token probabilities—perhaps utilizing perplexity-independent uncertainty metrics or dedicated verifier heads—to effectively identify when debate is necessary.

Future Work

To address the "Syntactic Determinism" failure observed in our GSM8K experiments, future work must prioritize creating a more robust confidence gating mechanism. We propose replacing the internal logit-based gate with a dedicated Reviewer agent trained to detect semantic inconsistencies independent of formatting. Furthermore, we aim to refine the heterogeneous debate panel by adopting agent roles from the X-MAS framework, specifically distinguishing between Solver and Reviewer agents, to further mitigate the echo chamber effect in complex reasoning tasks.

Societal Impact

These are language models, not oracles, and multi-agent debate does not guarantee objective truth. LLMs would almost always converge on a single answer, but not always a correct one. Particularly for highly ethical scenarios, such as judicial and medical decisions, mere consensus between different LLM's is not sufficient to make an ultimate decision when there are ethical or risk-related concerns. Therefore, all LLM answers should be subject to scrutiny by human auditors.

References

Balaji, P. G.; and Srinivasan, D. 2010. *An Introduction to Multi-Agent Systems*, 1–27. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-14435-6.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Dong, K. 2024. Large Language Model Applied in Multi-agent System: A Survey. *Applied and Computational Engineering*, 109: 9–16.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2024. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325.

Eo, S.; Moon, H.; Zi, E. H.; Park, C.; and Lim, H. 2025. Debate Only When Necessary: Adaptive Multiagent Collaboration for Efficient LLM Reasoning. arXiv:2504.05047.

Llama Team, A. . M. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Mistral AI Team. 2024. Mathstral 7B. https://mistral.ai/news/mathstral/. Hugging Face Model Card. Accessed: 2025-12-08.

Ye, R.; Liu, X.; Wu, Q.; Pang, X.; Yin, Z.; Bai, L.; and Chen, S. 2025. X-MAS: Towards Building Multi-Agent Systems with Heterogeneous LLMs. arXiv:2505.16997.

Zhang, H.; Cui, Z.; Chen, J.; Wang, X.; Zhang, Q.; Wang, Z.; Wu, D.; and Hu, S. 2025. Stop Overvaluing Multi-Agent Debate – We Must Rethink Evaluation and Embrace Model Heterogeneity. arXiv:2502.08788.

Individual Contributions

Eli Wiegman:

I selected the original Multi-Agent Debate paper used in the replication project, conducted early GitHub setup and API testing for replication. As we produced results for the replication, I assisted in generating visualizations using Python and Plotly, slide synthesis.

Wenjia Lu:

I contributed to the replication project codebase as well as ran benchmark evaluations for the LLMs. I also participated in effectively bridging our replication project to the extension, and heavily contributed to content on the research paper including but not limited to the abstract, intro, and backgrounds, as well as proofread and edited the research paper. I also contributed to our presentation slides.

Daniel Vega:

Transitioning raw data output from replication benchmarks and writing code that visualized this data in a parallel and efficient manner to the MAD paper visualizations. This process involved integrating JSON output into a Plotly framework that followed a similar front-facing interface used in the MAD research paper. Furthermore, my written contributions consisted of writing the Methodology section and integrating visuals, as well as writing the Future Work and Societal Impact sections of the Conclusion.

Ethan Justice:

I worked on selecting the DOWN paper to incorporate into the extension design, creating a generalized LLM wrapper class to enable the replication code and extension code remain the same regardless of model type or running location. I also implemented and ran the extension code, along with assisting with technical details for the paper for the extension portion.

Satyak Khare:

I was responsible for adapting the original Multi-Agent Debate codebase to support open-source inference, specifically refactoring the API wrappers to interface with vLLM. I managed the deployment and testing of the GPT-3.5-turbo and Llama-3.1 models on the University of Michigan's Great Lakes HPC cluster, troubleshooting environment and memory constraints. I also assisted with specific technical details for the writeup around the replication portion.

Christopher Zhou:

My main responsibility was being the project manager of the group. I organized team meetings to facilitate consistent progress of the project, and led the agenda for such meetings to ensure all sides from proposal, to codebase setup, to testing were represented. When our team ran into a major obstacle concerning the feasibility of our extension, I concretely identified how we could pivot our framework to a more reasonable solution while maintaining the distinct identity of the related works we covered.

Appendix

Agents/Rounds	MMLU	GSM8K	Biography
1/1	61%	78%	46%
3/2	63%	78%	45%

Table 5: Comparison of accuracy scores between the Single-Agent baseline (1/1) and the Multi-Agent setup (3 agents, 2 rounds) across three benchmarks for GPT-3.5-Turbo.

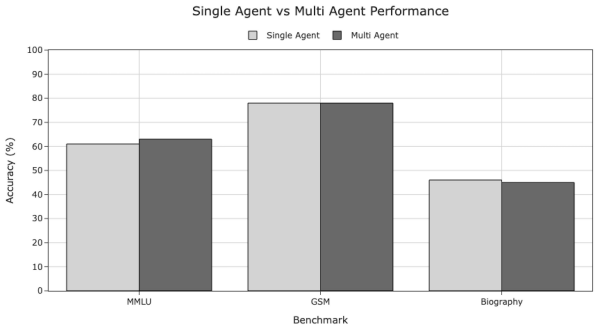

Figure 4: Replication with GPT-3.5 Turbo

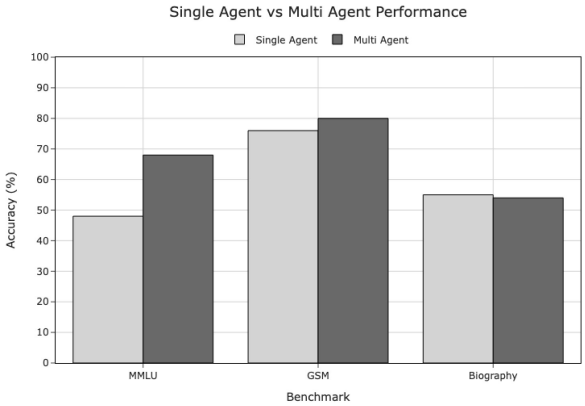

Figure 5: Replication with Meta-LLama-3.1-8B