Group 13

Building Efficient Heterogeneous Multi-Agent Debate Frameworks for Factual LLMs

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Ethan Justice, Satyak Khare, Daniel Vega, Wenjia Lu, Chris Zhou, Eli Wiegman

Motivation

- **Confusion & Hallucination**: Frontier LLMs still hallucinate, overstate confidence, and fail on multi-step logical reasoning.
- **One-shot thinking is fragile**: Suggestible models anchor to first reasoning path and rarely stray from first judgement.
- **Key Idea**: Treat reasoning as collaborative deliberation, not a single model response.

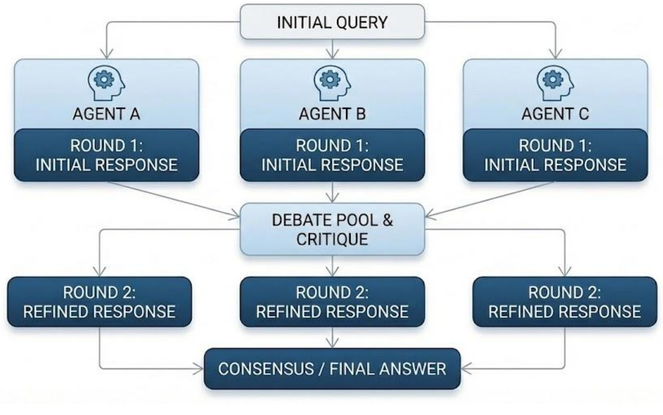
Research Question

Can combining confidence gating with heterogeneous debate reduce computational overhead (FLOPs) while maintaining or improving model accuracy?

Replication

Model Choice:
Llama-3.1-8B: Open source, fewer guardrails, raw results, enables human-like debate between agents

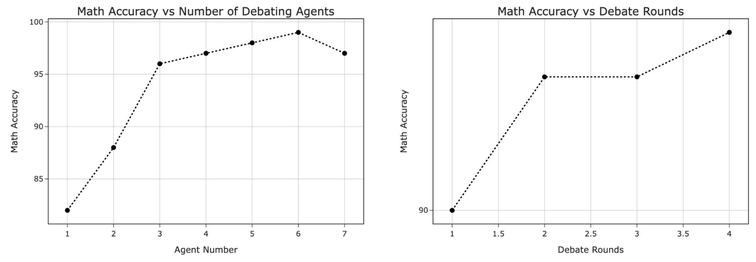
Methods:
Ran on **Single Agent** (1 agent - 1 round) & **Multi Agent** (3 agents - 2 rounds) on MMLU on GSM & Biographies benchmarks, initial query passed to each agent, each generate answer, other agents told to critique and modify their answer based on other answers. Repeats for n rounds, then final answer generated.



Replication Results

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

Here, we see an increase in MMLU and GSM results by the use of a multiagent system, highlighting the effectiveness of MAD for reducing reasoning based errors. However, we see a negligible decrease from single agent to MAD in biographies dataset revealing weaknesses in homogenous MAD for knowledge retrieval. The impact of different number of agents and rounds is seen below



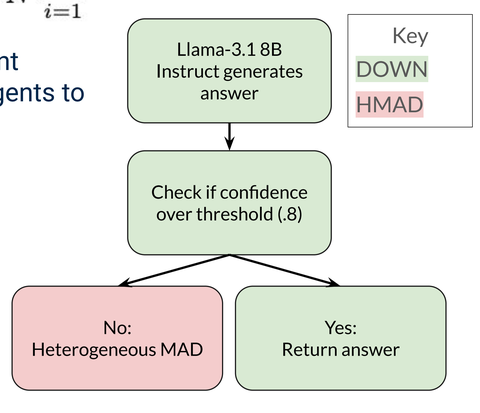
Extension

DOWN: Confidence Gating
To determine the initial LLMs confidence in its answer we used a confidence score defined as the average probability of each token, which is the average softmax score of the token logits as shown in the formula below. The model was considered confident if this score (c1) was above 0.8

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

Heterogeneous MAD:
We used 3 models with different strengths as heterogeneous agents to best cover blind spots.

Llama-3.1 8B Instruct	General Reasoning
DeepSeek R1 Distill Qwen 7B	Synthetic Reasoning Specialist
Mathstral 7B	Math Specialist



Extension Results

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**
,	1.000	Formatting
000	1.000	Formatting

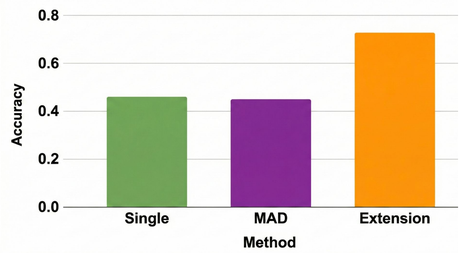


Table 1: Performance and computational cost by route.

Test	Route	Percent	Accuracy	Avg FLOPs
Biography	Gated	44%	0.7302	6.19 T
	Debated	56%	0.7308	75.31 T
	Overall	100%	0.7303	44.90 T
GSM	Gated	100%	0.3000	2.74 T
	Debated	0%	—	—
	Overall	100%	0.3000	2.74 T

Biography: The gate successfully routed "easy" queries to the single agent (12x compute reduction) and "hard" queries to debate. This selective prediction achieved an 18% accuracy gain (55% → 73%) with 40% less total compute, as HMAD solved complex queries the baseline MAD missed.
GSM: The confidence score failed due to Syntactic Determinism. Models were overconfident in the structure of the tokens (formatting, operators) rather than arithmetic correctness. Consequently, plausible-looking but incorrect answers bypassed the gate, preventing the necessary debate from triggering.

Conclusion and Future Work

The extension proves the potential effectiveness of this framework by achieving significant efficiency and accuracy gains with a 40% decrease in overall compute for the biography dataset, and an 18% accuracy increase. However the GSM benchmark highlights a limitations within the confidence scoring mechanism of overinflating the confidence score of heavily formulaic, and structured tokens, even if incorrect tokens fit the structure.

Future work entails optimizing the heterogeneous agents model selection and the addition of agent roles to improve performance, and the creation of a new generalizable confidence scoring mechanism that is able to determine confidence for all problem types.

Github + References

Final Paper PDF