Improving Multi-Agent Debate with Sparse Communication Topology

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, Eugene Ie

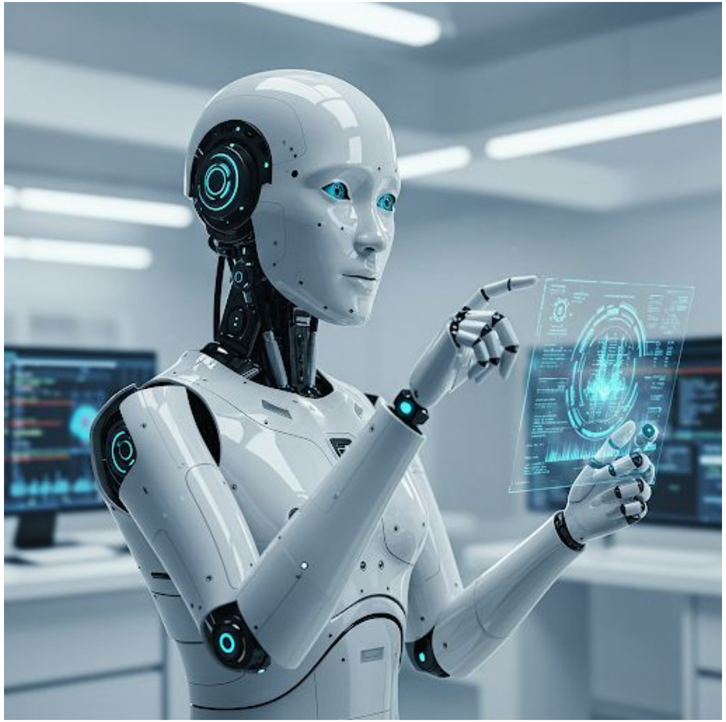
Workshop on Advancing LLM-Based Multi-Agent Collaboration
AAAI 2025 Workshop
Mar 4th, 2025

LLM-Based Agent

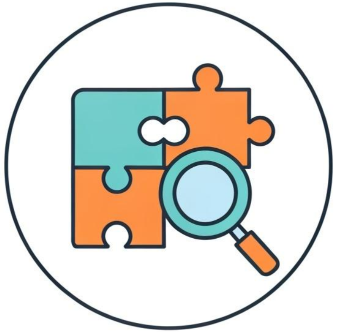
AI Agent - Autonomously perform tasks on behalf of a user or another system.

Key Capabilities

- Text & Multi-modal Understanding
- Reasoning & Decision Making
- Memory
- Tool use
- ...

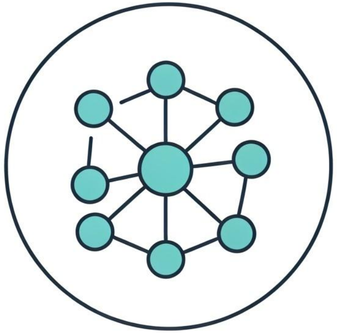


Multi-Agent



Problem Solving

Efficiency



Society of Mind

Multi-Agent Debate (MAD)

STEP 1: Initial Response Generation

Agents instantiated by LLMs generate solutions to a given question.

Multi-Agent Debate (MAD)

STEP 2: Multi-Agent Debate

Agent incorporates the responses of its connected peers from the previous round to debate using natural language for several rounds.

- ➔ Communication strategy
  - ◆ One-by-One
  - ◆ **Simultaneous-Talk**
  - ◆ Summarizer
- ➔ Communication topology
  - ◆ Fully-connected

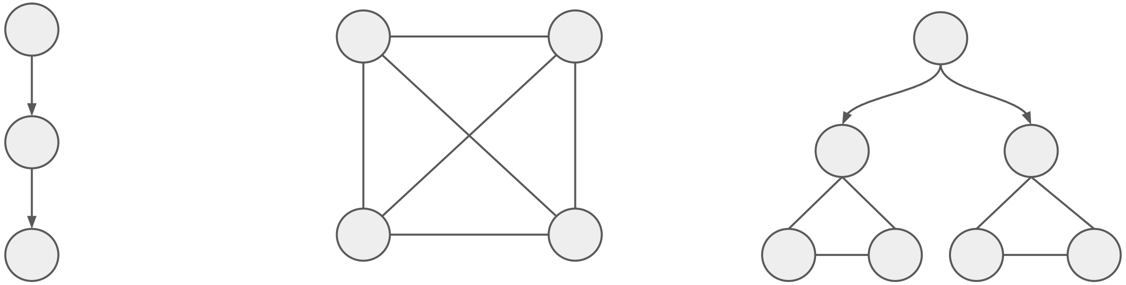
Multi-Agent Debate (MAD)

STEP 3: Reaching Consensus

Aggregate agents' responses to determine a consensus solution.

- Majority Vote
- LLM as a Judge

Communication Topology



- Communication topology can be complex, but is currently ill-studied in existing MAD.
    - Chain, tree, graph, hierarchical ...
    - Combination of above

Communication Topology Effect on Token Cost

Input token cost

debate rounds

user input length

debate history context

$$N \cdot R \cdot Q + N \cdot \boxed{C} \cdot H \cdot L$$

user input context

debate context

number of agents

agents connectivity

generation length

- Fully-connected: C = N - 1, leading to input token cost ~ O(N^2)

Isolating Sparse Communication Effects in MAD

Key Question: How does **sparsity** impact communication in a debate system?

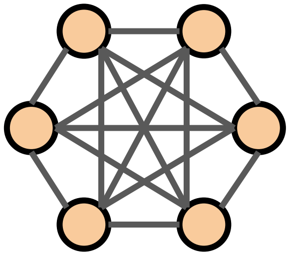
Analysis Approach: focus solely on the effect of sparsity, disentangle the impact of other factors

- agent roles
- specific topology patterns

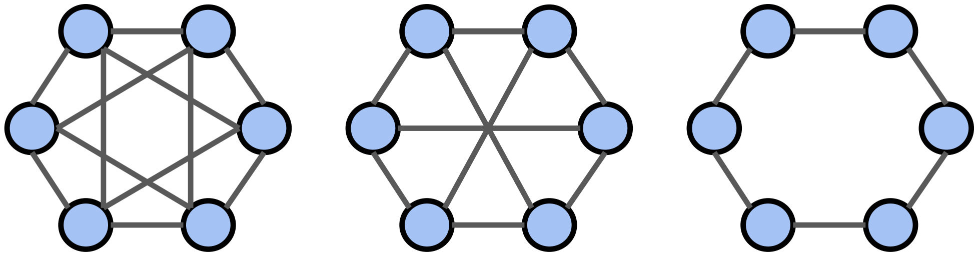
Analysis Approach

Focus: **Regular** Graph with **Homogeneous** LLM

Permutation Invariant: all agents are under the same position



Fully-Connected



Regular graph with various density: ⅘, ⅗, ⅖

$$D = \frac{2|\mathcal{E}|}{|\mathcal{V}|(|\mathcal{V}|-1)}$$

Analysis Approach

Focus: Regular Graph with Homogeneous LLM

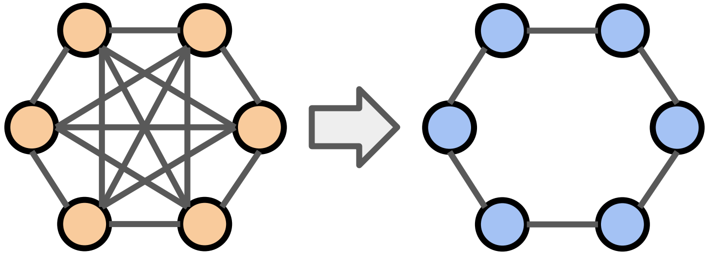
Connectivity Dynamics: deterministic v.s. randomized

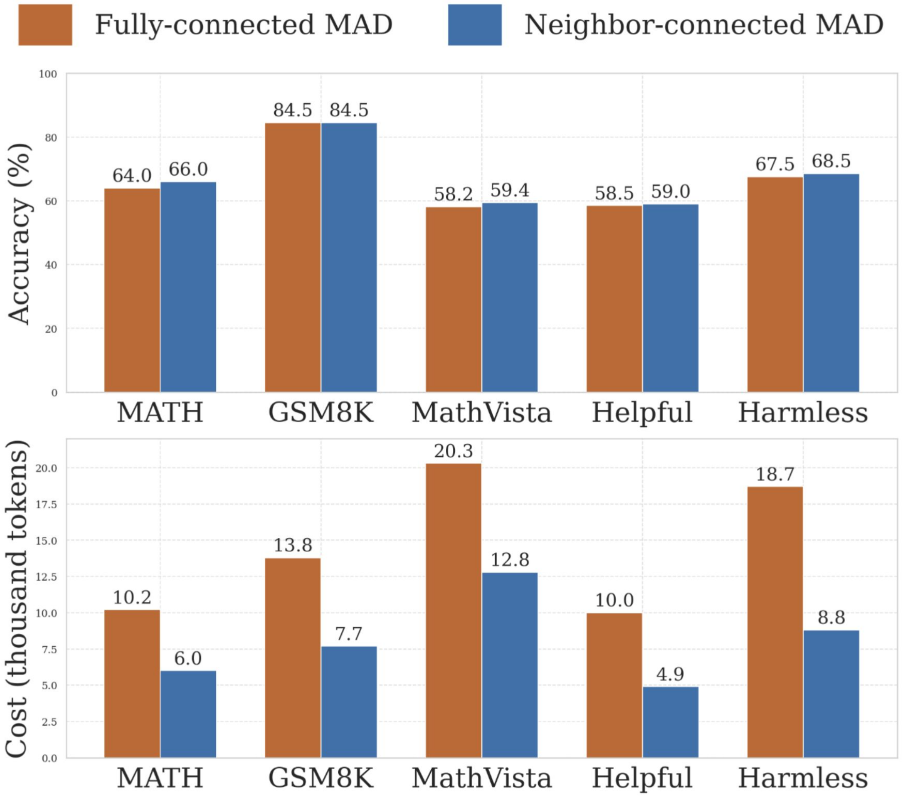
- **Deterministic**: topology is fixed during debate with density *D*.
- **Randomized**: the probability that a given agent sees any reference solution from previous round is *D*.

Experiments: Tasks

- Text Reasoning
  - GSM8K
  - MATH
- Multimodal Reasoning
  - MathVista
- Preference Modeling
  - Anthropic-Helpfulness
  - Anthropic-Harmlessness

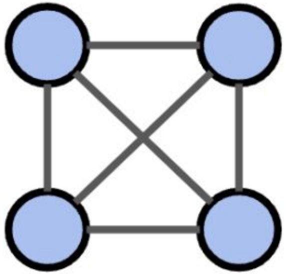
Performance of SparseMAD for N = 6



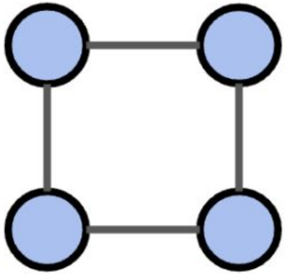


- On-par or slightly better quality (+1%)
- Significantly inference cost reduction (-40%)

SparseMAD for N = 4



Fully-Connected



Neighbor-Connected

Method	Accuracy	Cost
SC	81.0	-
$D = 1$	81.7 ± 0.9	baseline
$D = 2/3$	$\mathbf{82.7} \pm 1.2$	-25.6%

GSM8K task using the GPT-3.5 model.

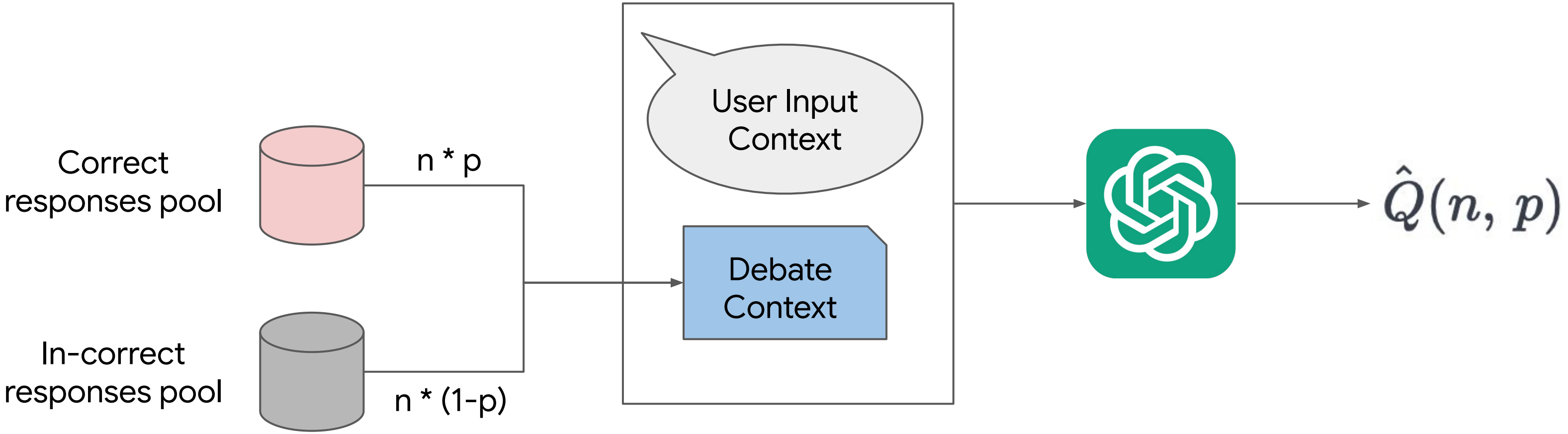
Randomized SparseMAD for N = 6

Method	Accuracy	Cost Saving
CoT	77.5 ± 4.2	-
SC	80.0	-
MAD ($D = 1$)	$\mathbf{84.5 \pm 1.5}$	baseline
ProbMAD ($D = 4/5$)	$\mathbf{84.5 \pm 0.7}$	-14.3%
ProbMAD ($D = 3/5$)	83.5 ± 0.7	-29.6%
ProbMAD ($D = 2/5$)	84.0 ± 1.7	-47.1%

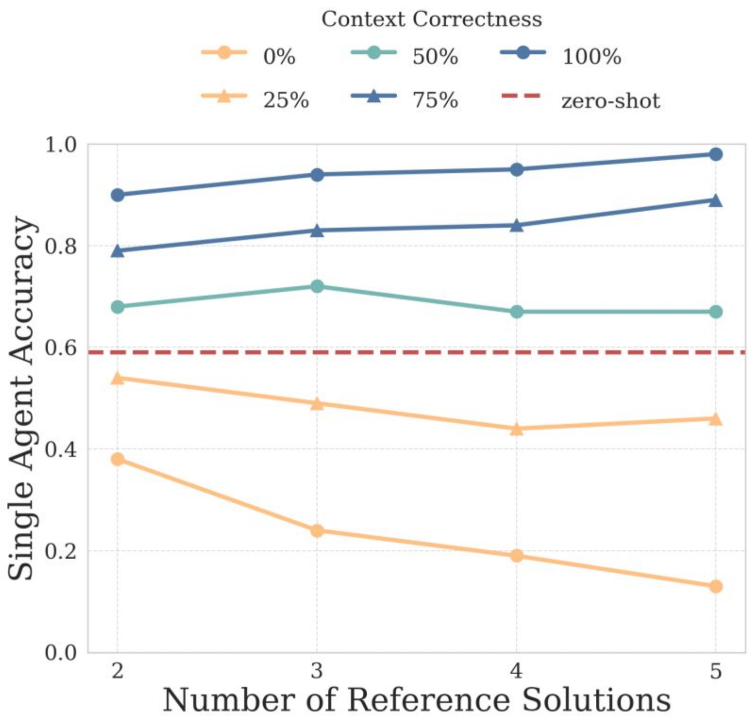
GSM8K task using the GPT-3.5 model.

Why Sparse Communication Topology Work?

Q(n, p): the probability that a single agent delivers correct answer, given **n** reference solutions where **p** percentage of them are correct.



Why Sparse Communication Topology Work?



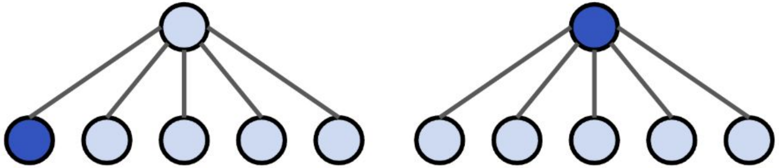
High context correctness: dense is better

Low context correctness: sparse is better

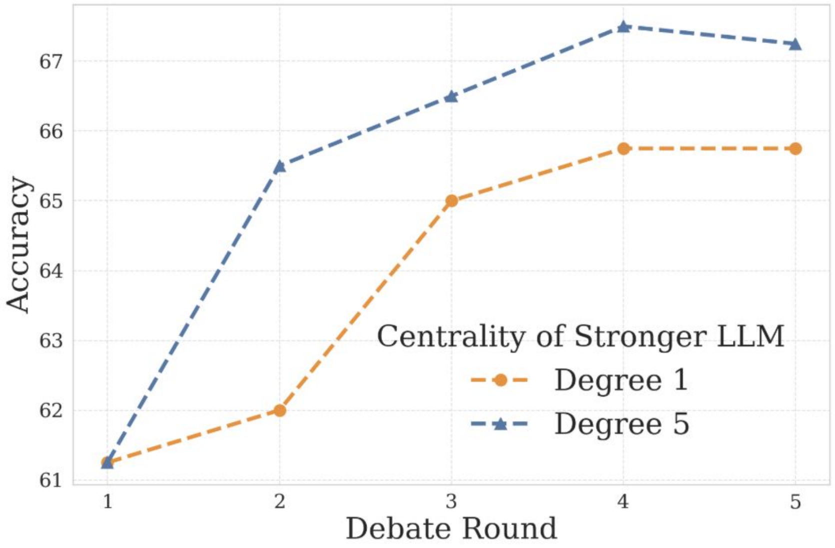
- When most agents do not provide correct answers, dense topology tends to mislead the agent into choosing incorrect answers.

Topology Design with Heterologous LLMs

Key Question: how to design the communication topology with different LLMs?



Isotropic Topology



- Put your **stronger** LLM on the **high-centrality** nodes

Conclusion

- Sparse communication topologies can improve the MAD performance significantly: **comparable** quality, **significantly reduce** costs.
- Extend the MAD framework to preference modeling tasks, demonstrating the benefits of MADs.
- Assigning stronger LLMs to **high-centrality** agent enhances overall performance.
- Present case-study insights that explain the effectiveness of sparse MADs.

Thank You!

Backup Slides

SparseMAD, N = 6, GSM8K

Method	Accuracy	Cost Saving
CoT	77.5 ± 4.2	-
SC	80.0	-
MAD ($D = 1$)	84.5 ± 1.5	baseline
MAD ($D = 4/5$)	83.5 ± 0.5	-12.7%
MAD ($D = 3/5$)	$\mathbf{86.5} \pm \mathbf{1.5}$	-29.1%
MAD ($D = 2/5$)	84.5 ± 0.8	-43.6%

Table 2: Comparison of accuracy and cost savings of MAD against baseline methods on the GSM8K dataset. All experiments were conducted using the GPT-3.5 model.

SparseMAD, N = 6, MATH

Method	Accuracy	Cost Saving
CoT	58.0 ± 2.0	-
SC	60.0	-
MAD ($D = 1$)	64.0 ± 1.4	baseline
MAD ($D = 4/5$)	$\mathbf{67.5 \pm 2.0}$	-14.6%
MAD ($D = 3/5$)	63.0 ± 1.8	-29.2%
MAD ($D = 2/5$)	66.0 ± 2.3	-41.5%

Table 1: Comparison of accuracy and cost savings of MAD against baseline methods on the MATH dataset. All experiments were conducted using the GPT-3.5 model.

SparseMAD, N = 6, MathVista

Method	Accuracy	Cost Saving
CoT	52.4 ± 2.6	-
SC	53.0	-
MAD ($D = 1$)	58.2 ± 1.5	baseline
MAD ($D = 4/5$)	57.8 ± 1.9	-9.1% (-11.5%)
MAD ($D = 3/5$)	55.4 ± 0.9	-20.0% (-24.7%)
MAD ($D = 2/5$)	$\mathbf{59.4} \pm \mathbf{0.6}$	-33.1% (-40.6%)

Table 3: Comparison of accuracy and cost savings of MAD against baseline methods on the MathVista dataset. All experiments were conducted using the GPT-4o model with the default temperature $T = 1$. The cost saving percentages in parenthesis are computed without multimodal inputs.

SparseMAD, N = 6, Anthropic-HH

Method	**GPT-3.5**		**Mistral 7B**	
	Accuracy	Cost Saving	Accuracy	Cost Saving
CoT	56.5 ± 3.1	-	60.8 ± 1.2	-
Self-Consistency	57.0	-	62.6	-
MAD ($D = 1$)	58.5 ± 1.7	baseline	65.5 ± 0.6	baseline
MAD ($D = 4/5$)	$\mathbf{59.0 \pm 1.8}$	-17.5%	65.6 ± 0.9	-18.3%
MAD ($D = 3/5$)	57.0 ± 1.6	-32.5%	64.6 ± 0.6	-35.2%
MAD ($D = 2/5$)	$\mathbf{59.0 \pm 1.4}$	-50.0%	$\mathbf{66.6 \pm 0.5}$	-53.5%

Table 4: AI labeler alignment accuracy and cost savings of MAD compared with baselines on the helpfulness dataset for GPT-3.5 and Mistral 7B models.

Method	**GPT-3.5**		**Mistral 7B**	
	Accuracy	Cost Saving	Accuracy	Cost Saving
CoT	66.0 ± 4.8	-	58.2 ± 2.0	-
Self-Consistency	67.0	-	60.0	-
MAD ($D = 1$)	67.5 ± 0.6	baseline	60.7 ± 0.3	baseline
MAD ($D = 4/5$)	67.0 ± 0.8	-17.3%	$\mathbf{62.2 \pm 0.2}$	-17.9%
MAD ($D = 3/5$)	67.5 ± 1.0	-34.7%	60.4 ± 0.4	-34.3%
MAD ($D = 2/5$)	$\mathbf{68.5 \pm 0.7}$	-53.3%	61.7 ± 0.2	-52.2%

Table 5: AI labeler alignment accuracy and cost savings of MAD compared with baselines on the harmlessness dataset for GPT-3.5 and Mistral 7B models.

Common types of agent behaviors in MAD

The Learner: "Considering the information from other agents, [...] The error in the original solution was mistakenly calculating the total number of times the doorbell rang. By correcting this, we find that ..."

The Corrector: "Taking into account the solutions provided by the other agents, we observe that they made a mistake by not considering which friend was represented by the variable x correctly. The first friend was incorrectly identified as the second friend. Using the correct identification and reasoning, ..."

The Arbitrator: "We see inconsistencies in the mentioned solutions. Let's correct it..."

The Gullible: "From the calculations provided, it seems the correct total number of doorbell rings should be <u>[wrong answer]</u>. \n\nThus, the total number of doorbell rings the doorbell made is <u>[wrong answer]</u>."