MAD-Community: Network of Multi-Agent Debate Communities

Brian Kim
The University of Texas at Austin
Department of Electrical and Computer Engineering
briankim31415@gmail.com

Abstract

This project explores a novel approach to enhancing Multi-Agent Debate (MAD) by implementing interconnected communities of large language model (LLM) agents. Each "MAD community" engages in structured debates with meta-prompting, culminating in a final answer judged and passed along to other connected communities in a directed acyclic network. Evaluating different network configurations demonstrated the importance of diversity and response integration for improving accuracy in answering graduate-level questions. Results highlight that optimized configurations enhance reasoning performance, offering insights into scalable and effective multi-agent systems for complex decision-making tasks.

1. Introduction & Motivation

Large Language Models (LLMs) have vastly improved their capabilities, especially in the past few years. A standout feature that has been on the rise is their ability to reason [1]. Typically, it is the case that a single LLM agent is asked a question and it gives a response in return. But with the advent of reasoning capabilities in LLMs, the same question could be asked and contributed to by multiple agents that reason and debate with one another before giving a final answer. This paved way for the development of Multi-Agent Debate (MAD). Multiple LLM agents responding to each other in an almost humane manner to recreate a real human discussion can provide the same benefits in that "two heads are better than one".

The main objective of this project is to extend MAD by creating multiple "MAD communities" and creating a network of these communities. Each community 'node' will perform MAD and pass along its responses to its connected communities. This project will answer the research question of if there are any network designs that result in more accurate answers compared to other designs on difficult graduate-level questions.

2. Prior Work

Du et al., 2023 [2] proposed a foundational design where agents exchange responses over multiple rounds, creating an evolving debate. While simple, this framework demonstrates the potential for LLMs to model argumentation and rebuttals in dynamic exchanges, laying the groundwork for more advanced debate structures.

Building on this, Liang et al., 2023 [3] introduced a more sophisticated framework that assigns distinct roles—affirmative debater, negative debater, and judge—to agents. Each agent operates based on specific meta-prompts aligned with its role, enhancing creativity and strategy in the debate. The judge evaluates the arguments and delivers a final decision, adding a structured resolution to the process. This role-based system expands on Du et al.'s [2] approach by introducing more nuanced interactions and decision-making, improving the ability of LLMs to engage in structured, multi-agent debates.

This project will continue building upon the frameworks developed up by both previous works through the combination of different MAD techniques from them in addition to the new proposed MAD community network. This is done to answer the research question of if different network configurations will improve the network's overall logic, knowledge, and reasoning performance. The exchange of responses in Du et al.'s [2] work will be replicated in the debates within each community of the proposed design. And Liang et al.'s [3] role assignment through the usage of meta-prompts will be implemented in the 'judge' role present in each community as well as the final output community node in the network.

3. Approach

The main research question that will be addressed is if having multiple layers of interconnected communities of LLM agents, each performing MAD, will improve the overall answer accuracy of the network on a series of difficult questions. To accomplish this, 3 different network configurations will be explored and evaluated.

Each network will consist of connections between multiple LLM ‘communities’. These communities contain multiple LLM agents, running GPT-4o mini, that will participate in MAD for a set number of rounds. These agents will be given meta-prompts that invoke disagreement and correction to encourage agents to deeply reason through their answers. After the debate has concluded, a ‘judge’ agent will select a final answer for that community. The community will then pass on its final answer and reasoning to its connected community(ies).

An example is shown in Figure 1, a community with 2 agents. Both agents receive the community’s input—to be further explained in the next paragraph. The lines represent flow of responses, consisting of the answer and reasoning. Agent 1 begins the debate and passes along its answer and reason to Agent 2. Agent 2 responds to Agent 1 and passes along its answer and reason to Agent 1. Agent 1 goes once more and passes its response to Agent 2. Both agents then pass of their responses to the Comunity Judge. The Community Judge chooses between one of the 2 responses it received and gives its output as the final response of the entire community.

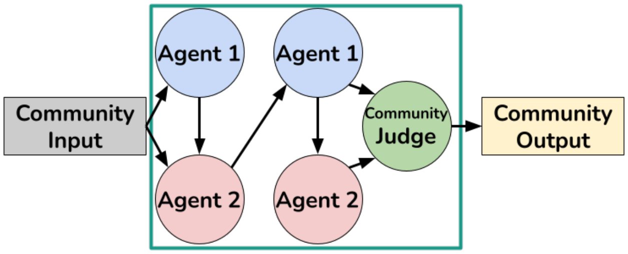

Figure 1. Message passing of a community with 2 agents and 2 rounds of debate

These networks are directed acyclic graphs, which means they have start and end nodes. The starting node(s) will be community(ies) that only debate over the given question. Each successive community after it will debate over the responses they receive. Eventually, these responses will propogate to the end node. This end node will be a single ‘judge’ agent that unilaterally chooses one of the responses it receives as the final answer for that network.

An example is given below in Figure 2. This network consists of 4 community nodes and 1 Network Judge. Each of the 4 communities can be broken down into the diagram shown Figure 1. The question is sent to Community 1 and Community 2. Community 1 prepares a response and immediately passes it off to the Network Judge. Community 2, however, passes its response off to Community 3 which passes off its response to Community 4. Community 4 finally passes its response to the Network Judge. The Network Judge then chooses either Community 1’s or Community 4’s response to output as the entire network’s final answer. This answer is then compared against the question’s correct answer from the database to measure accuracy.

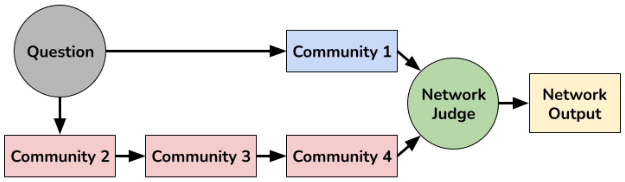

Figure 2. Short/long paths network design

Additionally, this network serves as the baseline for the experimentation on network design. The intuition behind this specific model was to compare the performance of a single MAD community (short path) versus multiple MAD communities passing off their responses to one another (long path). By comparing the accuracy of only Community 1 against the accuracy of Community 4, the linkage performance can be directly measured. And by comparing the accuracy of the short and long paths to the accuracy of the entire network, the benefits of having a network judge can be assessed.

To explore creative responses, the *temperature* parameter is also varied amongst communities. This parameter built into OpenAI’s API is a float where ”higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic” [4]. This enables one community to have more grounded and deterministic responses and the other to have more random and non-deterministic responses, empowering creativity within the network.

Initially, CosmosQA [5] was the original planned dataset to perform on. However, intial testing showed that even the cheapest ChatGPT models today greatly outperformed models used in the original paper, making performance evaluation much more difficult. As a result, the main dataset was changed to GPQA [6], a “graduate-level Google-proof Q&A” benchmark. These questions require much more knowledge in science domains such as biology, physics, and chemistry. Experts who have or are pursuing PhDs were able to achieve 65% accuracy and a GPT-4 based baseline reached an accuracy of 39%. Initial testing on the cheaper and smaller GPT4o-mini resulted in roughly 30% accuracy. This dataset would be able to show any improvements on a much larger scale than CosmosQA.

4. Results

4.1 Optimization of Parameters

Before testing various network designs, key parameters were optimized to improve overall performance. We will measure performance based on accuracy of the responses. That is, the network or community will be evaluated on if it provides the correct answer choice in its response. This will be represented as a percentage of questions it gets correct out of the total number of questions being tested on. Three primary adjustments were made: refining the system and user prompts for each agent and judge, adjusting temperature parameters, and exploring different configurations for the number of agents and rounds in Multi-Agent Debate (MAD) communities.

Each ChatGPT query involved a system prompt and a user prompt. The system prompt provided contextual behavioral guidance and task-specific instructions, while the user prompt contained the query to be addressed. In previous iterations, the system prompt included the question, and the user prompt focused on responses to other agents, following the methodology of Liang et al [3]. To improve clarity and precision, these prompts were restructured to place the question in the user prompt and provide more direct instructions to the agents. This adjustment significantly increased accuracy, as shown in Table 1. When tested on a set of 30 questions using the Short/Long Paths model with 2 agents and 2 rounds per community, the original prompt structure (Version 1) achieved 20% accuracy, while the updated prompts (Version 2) improved accuracy to 30%, representing a 10% increase.

Temperature parameters for agents and judges were also optimized. Version 1 utilized a temperature of 1.5 for agents and 0.5 for judges. These values were reduced to 0.7 and 0.3, respectively, in Version 2, leading to a substantial improvement in network accuracy from 30% to 50%. Based on these results, all subsequent experiments adopted the optimized temperature values of 0.7 for agents and 0.3 for judges.

Prompt Version	Agent Temps	Judge Temps	Network Accuracy
Version 1	1.5	0.5	20
Version 2	1.5	0.5	30
Version 2	0.7	0.3	50

Table 1. Prompt comparison

4.2 Comparing MAD Community Configurations

The next phase of testing evaluated the impact of different MAD community configurations on network accuracy. Specifically, the study compared whether increasing the number of agents per community or the number of debate rounds per community yielded better performance. Two configurations of the Short/Long Paths model were tested on 50 questions: one with 2 agents and 3 rounds and the other with 3 agents and 2 rounds. As summarized in Table 2, the 3 agents and 2 rounds configuration achieved a network accuracy of 54%, outperforming the 2 agents and 3 rounds configuration, which achieved 38% accuracy. This improvement suggests that increasing the number of agents, and therefore the diversity of viewpoints, is more beneficial than increasing the number of debate rounds.

Further analysis revealed that while the accuracy of the short path (Community 1) was similar between the two configurations, the accuracy of the long path (Communities 2, 3, and 4) differed significantly (34% in 2 agents and 3 rounds versus 50% in 3 agents and 2 rounds). This discrepancy may be attributed to the greater number of unique viewpoints in the 3 agents and 2 rounds configuration, where each community receives responses from three agents and one prior community, compared to two agents and one prior community in the 2 agents and 3 rounds configuration.

Community Config	Short Path Accuracy	Long Path Accuracy	Network Accuracy
2 agents, 3 rounds	42%	34%	38%
3 agents, 2 rounds	46%	50%	54%

Table 2. Community Configuration

4.3 Evaluation of Network Designs

Using the optimized parameters and configurations, three network designs—Short/Long Paths, Feed Forward, and Cascade—were tested on 50 questions. Each design was evaluated using both the 2 agents and 2 rounds and 3 agents and 2 rounds configurations.

1. **Short/Long Paths Network**
   In the Short/Long Paths network, short path accuracy was determined by Community 1, while long path accuracy was based on Community 4 (Figure 2). As seen in the results shown in Table 3, the long path accuracy remained consistent at 50% for both configurations. However, the short path accuracy improved by 8% in

the 3 agents and 2 rounds configuration compared to the 2 agents and 2 rounds configuration, highlighting the benefits of incorporating additional agents.

2. **Feed Forward Network**
   The Feed Forward network consisted of two fully connected layers, where each community in the first layer sent responses to all communities in the second layer, which then passed their responses to a final judge (Figure 3). Contrary to expectations, the average accuracy of second-layer communities was lower than that of first-layer communities by 4.6% in the 2 agents and 2 rounds configuration and 4.7% in the 3 agents and 2 rounds configuration (Table 4). This trend resulted in the Feed Forward network being less accurate overall compared to the Short/Long Paths network, potentially due to information dilution when multiple inputs were fed into a single community.

3. **Cascade Network**
   The Cascade network featured a triangular structure, with a decreasing number of communities per layer (Figure 4). Communities in each layer sent responses to one or two communities in the next layer, culminating in a single judge node. Interestingly, accuracy remained relatively stable across layers, with only a 3.5% decrease between the first and third layers in the 3 agents and 2 rounds configuration as seen in Table 5. Similar to the Feed Forward network, a general decline in accuracy across successive layers was observed, possibly due to the increased complexity of integrating multiple inputs. This trend was not evident in the Short/Long Paths network, suggesting that excessive aggregation of community responses may hinder performance.

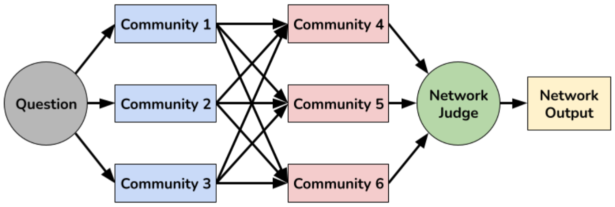

Figure 3. Feed Forward network design

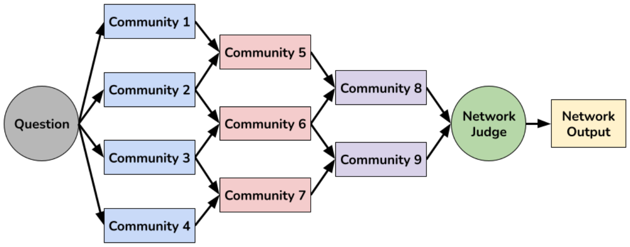

Figure 4. Cascade network design

Community Config	Short Path Accuracy	Long Path Accuracy	Network Accuracy
2 agents, 2 rounds	42%	50%	46%
3 agents, 2 rounds	46%	50%	54%

Table 3. Short/Long Paths network results

5. Conclusion & Future Work

This project aims at exploring Multi-Agent Debate (MAD) and implementing it within communities of LLM agents and constructing a network of these communities. The study demonstrated that prompt structure, temperature settings, and MAD community configurations significantly impact the performance of networks in Multi-Agent Debate scenarios. The optimized system achieved a 20% improvement in accuracy by refining prompts and temperature values. Among the three network designs, the Short/Long Paths network consistently outperformed the Feed Forward and Cascade networks, with its long path accuracy unaffected by successive layers. The findings highlight the importance of balancing viewpoint diversity and response integration to maximize accuracy in complex decision-making systems.

Immediate future work include testing on more questions and fine-tuning for optimal parameters such as temperature and prompt engineering. Beyond these, I believe exploring network designs that implmement some sort of recursive or cyclic loops would be interesting. A design idea that was not able to be tested was the proposed Hopfield network model. This model would consist of multiple nodes that are all fully connected with each other and pass responses to one another until all nodes converge onto the same answer choice. However, efficiency and cost would also be other metrics to consider when taking this approach. As a result, future works studying the economy of developing and run-

Community Configuration	Layer 1 Accuracy	Layer 2 Accuracy	Network Accuracy
2 agents, 2 rounds	43.3%	38.7%	36%
3 agents, 2 rounds	48%	43.3%	42%

Table 4. Feed Forward network results

Community Configuration	Layer 1 Accuracy	Layer 2 Accuracy	Layer 3 Accuracy	Network Accuracy
2 agents, 2 rounds	46.5%	46.7%	45%	44%
3 agents, 2 rounds	47.5%	44.7%	44%	44%

Table 5. Cascade network results

ning efficient networks could also be a very fruitful route.

6. Lessons Learned

Through this project, I have gained valuable experience working with the ChatGPT API. This has greatly enhanced my skills in integrating LLMs into real-world applications. Additionally, I was able to practice effective prompt engineering techniques, enabling me to write precise and contextually appropriate prompts that maximize the performance of LLMs. Beyond these technical skills, this project deepened my understanding of network structures and their underlying principles. I was able to discover the profound impact that network design has on model behavior and outcomes through the MAD communities. This holistic perspective has not only broadened my technical expertise, but also strengthened my appreciation for the interplay between model design and practical implementation in AI systems.

References

[1] Boshi Wang, Xiang Yue, and Huan Sun. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of EMNLP*, 2023.

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

[3] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[4] OpenAI API Reference, https://platform.openai.com/docs/api-reference.

[5] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*, 2019.

[6] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

TEXAS
The University of Texas at Austin

Brian Kim (bmk994) B(

Course Evaluations Fall 2024

Hi Brian Kim (bmk994), you have been invited to complete evaluations for the following courses.

C S 394N - NEURAL NETWORKS (50840, 55560)

● Open   Ends on: **2024-12-09**

Start now

ECE 381K - 20-MCHN LRNG REAL WORLD NETWRK (17695)

● Completed   Ends on: **2024-12-09**