MAD-Community

Brian Kim

ECE 381K - Poster Session
Dec 5, 2024

1. Introduction

Current LLM Reasoning Capabilities

- Reasoning capabilities of LLMs have increased greatly over the past few years.
  - On MMLU benchmark, SOTA LLM performance has increased nearly 200% from 2019 to 2024.

Multi-Agent Debate

- The Multi-Agent Debate (MAD) framework has multiple LLM agents communicate with each other and 'debate'.
- Agents critique and improve each others' responses to arrive at the correct answer.

2. Problem description

MAD Communities

- Create multiple MAD 'communities' and connect them in a directed acyclic graph network.
- Use output response of a community as the input of its connected communities.

GPQA

- A graduate-level Google-Proof Q&A benchmark written by domain experts in biology, physics, and chemistry.
  - Strongest GPT-4 based baseline achieved 39% accuracy.

3. Approach

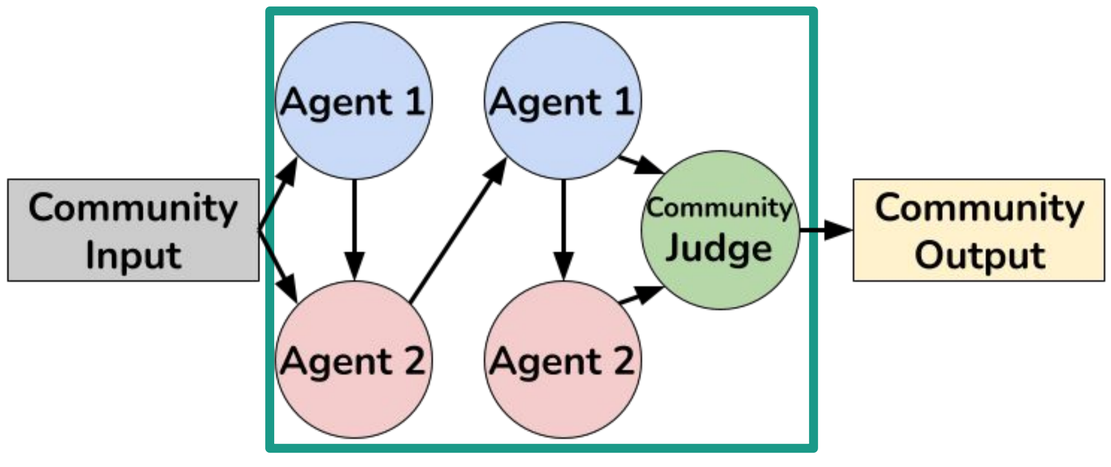

Fig 1. Community node with 2 agents, 2 rounds

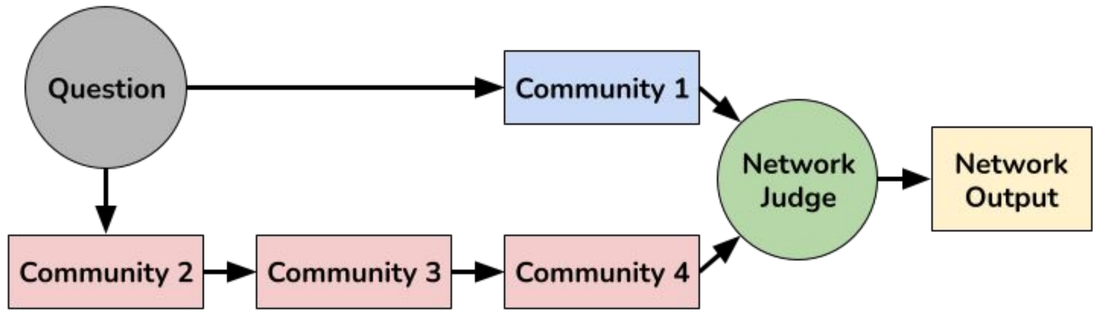

Fig 2. Short/Long Paths Model

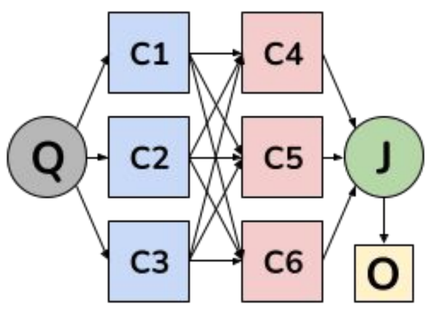

Fig 3. Feed Forward Model

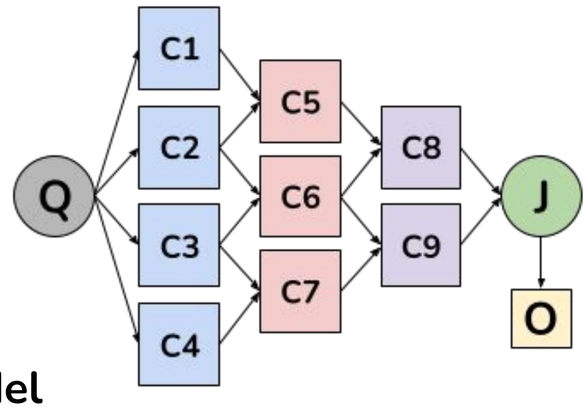

Fig 4. Cascade Model

4. Results

Plot

Model	30 Questions		50 Questions	
	Agents' Score	Network Score	Agents' Score	Network Score
Short/Long Paths	40%	33.3%	46%	***46%***
Feed Forward	***57%***	***43.4%***	***55%***	36%
Cascade	42%	36.7%	***55%***	44%

Summary

- Feed forward model saw greatest delta from agent score to network score.
  - May be caused by the network's fully connected design since first 3 communities averaged 43.3% correct but last 3 communities averaged 38.7% correct on 50 questions.
- On average, short/long paths saw most consistent performance between agent to network score.
- Comparable performance between MAD-Community GPT4o-mini agents and GPQA's strongest GPT-4 based baseline while being significantly cheaper per query ($0.15/1M tokens vs $30/1M tokens).