Multi-Turn LLM Systems for Diagnostic Decision-Making: Considerations, Biases, and Challenges

Benjamin Liu[1], Sejong Kim[3], Drona Thoka[3], Varun Puttagunta[3], Kaylin Sheng[3], Mark Li[4], Kiran Nijjer[2,3], Adnan Ahmed[1], Thi Uyen Hanh Le[5], Sai Chidvilas Gudiboina[7], Ali Ugur[6], Sunishchal Dev[3], Kevin Zhu[3]

[1]York University; [2]Stanford School of Medicine; [3]Algoverse AI Research; [4]Carnegie Mellon University; [5]Boston University; [6]Rutgers University, [7]Birla Institute of Technology and Science

Introduction/Background

- LLM-MAS simulates cooperation with agent roles such as: doctors, patients, test-providers, experts, and evaluators. Each agent takes roles in **questioning, analyzing, and diagnosing** with **multi-turn dialogue.**
- The study uses **GPT 4.1** under the **Agent Clinic Framework** to evaluate **architecture**, **information access**, and **social biases** to understand the critical components of the system.
- We evaluate how these systems **respond to challenges** that human medical teams face, such as **limited patient-information** and **limited dialogue length**.
- We further examine how multi-turn interactions can amplify or suppress **emerging biases**, highlighting **safety considerations** for **real-world usage**.

Objectives/Aims

- The study aims to determine the most and least **important information** in each of the modular parts of the LLM-MAS.
- The results help to create **novel architectures** for use in clinical cases. By modeling real world workflows, the AI can **significantly boost** its **accuracy, efficiency, and communicability**.
- The research can be applied to **MAS in other domains,** and prove which information is **most valuable** to an AI team.
- Evaluate how multi-turn interactions introduce or exacerbate **demographic and social biases** in diagnostic reasoning, and identify which **elements most strongly contribute** to those **inequalities**.

Methods

- The AgentClinic benchmark was used to test how **effective LLM-MAS** are in diagnostic reasoning. The framework simulates multi-turn clinical dialogues among agents to study how **information** is **gathered**, **interpreted**, and **confirmed** under **incomplete data and time limits**, using GPT-4.1.
- **Agent Roles:**
  - **Doctor**: Interact and provide Diagnosis
  - **Patient**: Answer doctor questions and provide background
  - **Specialist**: Consult with doctor and assist diagnosis with its expertise
  - **Measurement**: Supply the results to requested medical tests
  - **Moderator**: Determine if the doctor reached a proper diagnosis

Figure 11: Adversarial Agent Interaction Workflow

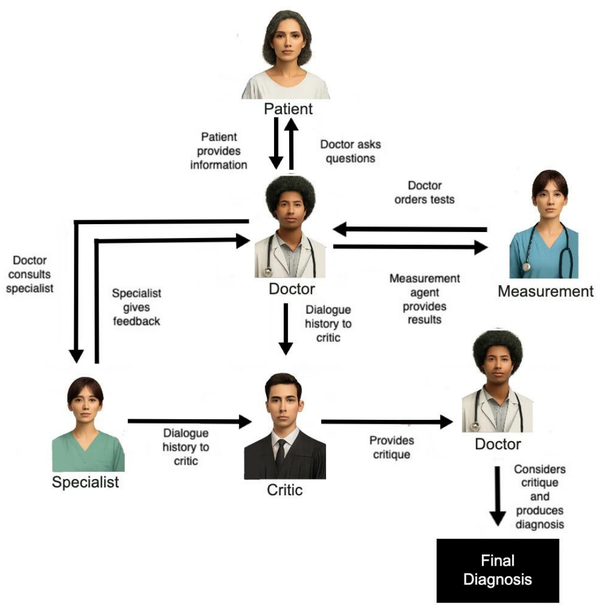

Figure 11: Flow chart of adversarial agent architecture interactions

Figure 12: Dynamic Agent Interaction Workflow

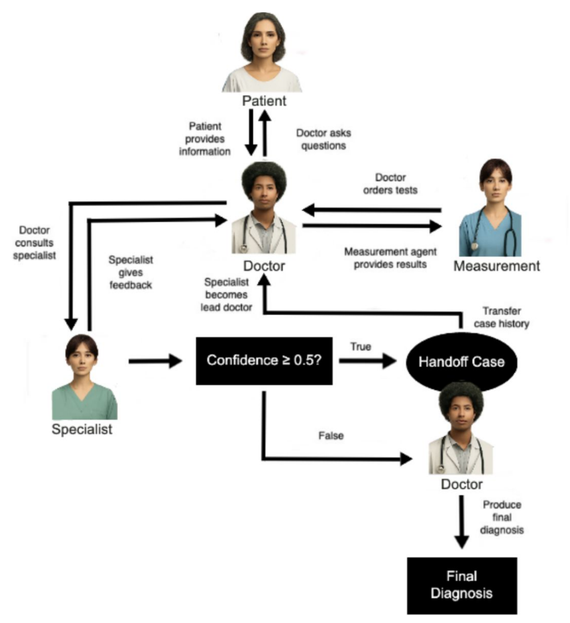

Figure 12: Flow chart of dynamic agent architecture interactions

Results

Table 1: Doctor Agent under various modes of **patient information access**

Information available	Diagnostic accuracy Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)
Chief complaint	22.00	35.33	44.00	47.33
+ Symptoms	48.67	58.67	65.33	67.33
+ History	54.67	70.67	76.67	76.67
+ Demographics	56.00	70.67	75.33	77.33

Diagnostic process metrics Avg dx considered	Avg tests	Efficiency / Information quality Avg emb	Avg best emb	Avg info den
11.04	3.04	0.41	0.83	0.44
10.17	4.19	0.46	0.84	0.46
9.01	4.64	0.47	0.84	0.46
9.44	4.19	0.46	0.81	0.46

Figure 1: Doctor-patient **turn length** accuracies by top-K

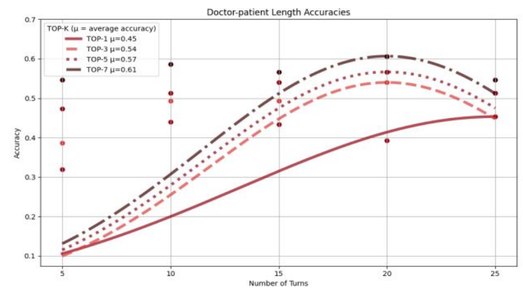

Table 2: Measurement Agent under various modes of **patient information access**

Tests available	Diagnostic accuracy Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Diagnostic process metrics Avg dx considered	Avg tests	Efficiency / Information quality Avg emb	Avg best emb	Avg info den
EKG	52.60	64.67	70.67	72.67	7.54	0.33	0.58	0.74	0.58
+ Blood tests	50.00	67.33	71.33	72.67	7.39	0.33	0.59	0.75	0.59
+ Physical	50.00	65.33	68.67	70.00	7.63	0.25	0.58	0.73	0.58
+ Vital	52.00	64.67	69.33	75.33	7.48	0.31	0.58	0.76	0.58

Table 4: Doctor specialist consultation under **various length constraints**

# of turns	Diagnostic accuracy Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)
5 Turns	34.67	42.00	44.67	52.33
10 Turns	39.33	45.33	49.33	52.00
15 Turns	36.00	46.67	47.33	54.67
20 Turns	41.33	47.33	51.33	56.67
25 Turns	38.00	41.67	48.00	54.00

Diagnostic process metrics Avg dx considered	Avg tests	Efficiency / Information quality Avg emb	Avg best emb	Avg info den
3.84	0.23	0.42	0.69	0.50
4.09	0.31	0.42	0.71	0.51
4.73	0.18	0.43	0.68	0.51
4.41	0.29	0.43	0.68	0.51
4.37	0.19	0.43	0.68	0.50

Figure 2: Doctor-specialist **turn length** accuracies by top-K

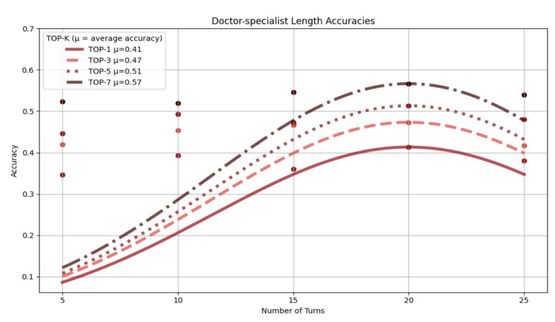

Table 6: **Various architectures** to model clinical workflow

Name	Diagnostic accuracy Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Diagnostic process metrics Avg dx considered	Avg tests	Efficiency / Information quality Avg emb	Avg best emb	Avg info den
Hierarchical	32.00	40.89	48.10	49.99	9.98	5.03	0.41	0.72	0.28
Redundant	45.66	48.00	52.10	54.39	10.19	7.53	0.47	0.82	0.24
Adversarial	64.67	76.00	80.67	82.67	10.13	0.46	0.46	0.84	0.50
Dynamic	53.33	64.00	67.33	74.00	10.11	0.00	0.44	0.74	1.00

Limitations

- **MedQA** is the only dataset we used for our experiments.
- All experiments were run only using **GPT-4.1**.
- Only **150** scenarios were run for each experiment.
- **4 architecture types were investigated**, covering a small subset of possible multi-agent reasoning structures.
- The metrics measured offer limited insight into **error propagation**, and **adaptability over extended dialogues**.
- The simulated interactions cannot capture **patient variability**, **clinical noise**, or the **unpredictability of human dialogue**.

Conclusions

- The **design of coordination strategies** and **information flow** is **essential** for achieving **reliable multi-turn reasoning**, particularly in scenarios where the number of **interactions is limited and data is incomplete**.
- **Architecture proved to be crucial**, showing that diverse reasoning and adaptive roles **improve diagnostic metrics**. In specific, the **Adversarial** and **Dynamic** architectures showed the most improvement.
- Fairness studies revealed persistent biases, particularly between **genders and lifestyle subgroups**, indicating that cooperative reasoning in LLMs does not fully mitigate the negative effects of bias.

Future Directions

- Expand beyond the **MedQA dataset** and run more scenarios for each experiment to maximize representativeness**.**
- Evaluate the behavior of **multiple LLM families** beyond the GPT family.
- Include more architectures, expanding beyond the four investigated in this experiment.
- Investigate **reinforcement learning** for sparse-reward environments.
- Design **richer multi-turn evaluation paradigms** to study agent coordination, and performance degradation over extended interactions.
- Understand how **biases** may **affect certain architectures** in particular.
- Analyze how different **architectural patterns** interact with demographic bias to identify configurations that **mitigate or exaggerate disparities**.
- Consider **computational cost metrics** and **statistical significance** for our variables.

Acknowledgements

- Thank you to **Algoverse AI Research** for their collaboration.

References