Adaptive Multi-Agent Diagnostic Reasoning: A MAS-GPT Inspired Framework for Clinical Decision Support

Tad Sligar
Auburn University
tms0073@auburn.edu

Abstract

We propose an adaptive multi-agent diagnostic reasoning framework inspired by **MAS-GPT**, designed to model collaborative clinical decision making. A generalist agent (Emergency Medicine, Pediatrics, or Family/Internal Medicine) performs triage and enumerates a predefined catalog of medical and surgical specialties before selecting the Top-5 most relevant consultants. This ensures interpretability, prevents hallucinated roles, and mirrors real-world multidisciplinary collaboration. Each specialist provides structured diagnostic reasoning, which the generalist aggregates into a final decision. This one-month project will build a prototype using the MedQA-USMLE dataset and qualitatively evaluate reasoning interpretability and system feasibility. The outcome will form the foundation for a future publication exploring adaptive agent generation for clinical reasoning.

1 Introduction

Clinical reasoning often requires synthesizing insights from multiple specialties—something large language models (LLMs) struggle to emulate. Traditional LLMs reason in a single forward pass, often skipping intermediate logic or hallucinating expertise outside their training domain. Multi-agent reasoning frameworks, such as **MAS-GPT**, demonstrate that decomposing complex reasoning into structured, collaborative steps can enhance both transparency and factual reliability.

This project explores whether a dynamically generated network of medical specialist agents, coordinated by a generalist "planner," can produce more interpretable and accurate diagnostic reasoning. The framework will be implemented and evaluated on a subset of the MedQA-USMLE benchmark, focusing on qualitative improvements rather than full-scale performance.

Research Question: Can an LLM-driven, dynamically generated team of specialist agents improve the interpretability and reliability of clinical reasoning compared to a single-model or static multi-agent approach?

2 Background and Related Work

Tree of Thoughts (ToT) [3] introduced structured reasoning trees, encouraging deliberate step-wise exploration. **AgentVerse** and **AutoGen** [4] extended this concept by enabling multiple LLMs to communicate through predefined roles and dialogue protocols. However, these systems rely on fixed, human-defined architectures.

MAS-GPT [1] advanced the field by allowing an LLM to dynamically generate both the agent definitions and the communication graph for each task. This adaptivity enables context-specific reasoning structures, but at the cost of interpretability and consistency.

Preprint. Under review.

In the medical domain, models such as **MedPaLM** [5], **BioGPT** [6], and **PMC-LLaMA** [7] demonstrate impressive factual accuracy but lack modular reasoning. They treat diagnosis as single-prompt question answering, without simulating real clinical workflows. Our proposed system combines MAS-GPT's adaptivity with clinical structure to achieve both flexibility and interpretability.

3 Methodology

This section outlines the core components of the proposed system: the planner, executor, safeguards, and baseline comparisons. The project focuses on reproducibility, interpretability, and modularity over scale.

3.1 Planner: Generalist Triage and Specialist Selection

A generalist agent serves as the "planner," determining the reasoning pathway based on case context. The generalist type is chosen by triage rules:

- **Emergency Medicine** – selected for unstable vital signs or red-flag findings.
- **Pediatrics** – selected if the patient is aged 17 or below.
- **Family/Internal Medicine** – default for adult, non-urgent cases.

Once triaged, the generalist enumerates a predefined catalog of specialties (medical, surgical, and subspecialty). Each is scored on:

1. *Relevance* – overlap with presenting complaint.
2. *Coverage Gain* – diversity of hypotheses added.
3. *Urgency Alignment* – likelihood of immediate risk.
4. *Procedural Signal* – relevance of surgical or procedural expertise.

The Top-5 specialties are then selected for consultation, ensuring comprehensive coverage while limiting agent proliferation.

3.2 Executor: Specialist Consultation and Aggregation

Each selected specialist acts as a focused reasoning module, generating a structured JSON output:

```
{
  "specialty": "Cardiology",
  "differential": [
    {"dx": "NSTEMI", "p": 0.82,
     "for": ["exertional chest pain", "radiation to jaw"],
     "against": ["pain reproducible by palpation"]}
  ]
}
```

The generalist aggregates these outputs by weighting evidence consistency and probability across specialists, producing an ordered differential diagnosis and a concise justification paragraph.

3.3 Safeguards and Constraints

To maintain control and reproducibility:

- A maximum of five specialist agents may be created per query.
- Temperature is fixed at 0.3 to encourage determinism.
- Circular communication between agents is prohibited.
- The specialty catalog is static to prevent hallucinated agents.

3.4 Baselines

To contextualize the system's performance, we will compare against:

- A single-LLM Chain-of-Thought baseline.
- A fixed four-agent pipeline (Planner, Specialist, Reviewer, Aggregator).
- A debate-style dual-agent reasoning model.

4 Evaluation Plan

The evaluation emphasizes interpretability, reasoning structure, and feasibility within the one-month project window.

4.1 Quantitative Metrics

- **Accuracy:** Match between predicted and ground-truth MedQA diagnosis.
- **Cost:** Average tokens or agents per query.
- **Latency:** Average inference time per complete reasoning chain.

4.2 Qualitative Metrics

- **Interpretability:** Human-readable, logically connected reasoning traces.
- **Reliability:** Consistency of specialty selection across repeated trials.
- **Transparency:** Absence of hallucinated specialists or unjustified steps.

4.3 Experimental Setup

A subset of 500 questions from the MedQA-USMLE dataset will be used. The LLM backbone (e.g., GPT-4 or Claude) will be accessed through API calls via a Python orchestration layer, enabling reproducible runs and structured output validation.

5 Expected Outcomes

We expect qualitative improvements in reasoning structure and factual reliability compared to the single-LLM baseline. The adaptive planner is hypothesized to:

- Improve interpretability through modularized specialist reasoning.
- Reduce irrelevant reasoning paths and hallucinations.
- Produce 10–20% relative improvement in top-1 diagnostic accuracy.
- Offer better traceability for human reviewers.

The project is primarily exploratory; statistical significance is not expected within the short timeframe, but results will hopefully guide a future extended study.

6 Project Timeline and Deliverables

Week 1: Literature review, data familiarization, and prompt design.

Week 2: Implement triage planner and specialty scoring logic.

Week 3: Develop specialist and aggregator modules; test JSON consistency.

Week 4: Evaluate MedQA subset, perform qualitative analysis, and write report.

Deliverables:

- Functional prototype (Python orchestration with LLM API).
- Example reasoning traces and annotated MedQA results.
- Final written report and oral presentation.

7 Discussion and Future Work

Challenges include ensuring deterministic reasoning under stochastic model outputs and validating reasoning correctness without clinician oversight. Future work will explore:

- Integration of retrieval-augmented generation using PubMed abstracts.
- Multi-turn conversational refinement between agents.
- Quantitative scaling to full MedQA and MMLU-Clinical benchmarks.
- Potential submission to NeurIPS LLM Reasoning Workshop or AMIA 2025.

8 Broader Impact

The modular approach improves transparency and allows auditing of LLM reasoning, which could inform future safe-use frameworks for medical AI systems. Ethical safeguards include restricted data sources, no patient information, and interpretability over automation.

9 Conclusion

This proposal presents an interpretable, adaptive multi-agent reasoning system for clinical question answering. By combining generalist triage, Top-5 specialist consultation, and structured aggregation, it models the collaborative logic of real clinical practice. The resulting prototype will serve as a proof-of-concept for an eventual publishable research effort on adaptive multi-agent reasoning in medicine.

References

[1] Guo et al., "MAS-GPT: Multi-Agent System with Generated Programs," 2024.

[2] Jin et al., "What Disease Does This Patient Have? MedQA-USMLE," 2020.

[3] Yao et al., "Tree of Thoughts: Deliberate Problem Solving with Large Language Models," 2023.

[4] Wu et al., "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," 2023.

[5] Singhal et al., "MedPaLM: Large Language Models for Medical Question Answering," 2023.

[6] Luo et al., "BioGPT: Generative Pre-trained Transformer for Biomedical Text," 2022.

[7] Wu et al., "PMC-LLaMA: Towards Open Medical Large Language Models," 2023.