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

Sejong Kim
sejong@ohs.stanford.edu

Drona Thoka
drona_thoka_student_@algoverseairesearch.org

Varun Puttagunta
varunputtagunta@my.unt.edu

Kaylin Sheng
kaylin_sheng_student@algoverseairearch.org

Mark Li
markli@andrew.cmu.edu

Kiran Nijjer
kirannijjer09@gmail.com

Adnan Ahmed
adahmed2@stanford.edu

Thi Uyen Hanh Le
katele2277@gmail.com

Sai Chidvilas Gudiboina
f20221212@goa.bits-pilani.ac.in

Ali Ugur
agu6@scarletmail.rutgers.edu

Kevin Zhu
kevin@algoverseairesearch.org

Abstract

This study investigates the systemic limitations and architectural design trade-offs of Large Language Model multi-agent systems (LLM-MAS) for clinical decision support, focusing on how agent collaboration and architectural choices influence reasoning in complex medical problems. We examined the effects of changes in agent roles, interaction protocols, and architecture on diagnostic accuracy and reasoning through targeted ablation studies with the AgentClinic framework. Reflecting the time-sensitive and uncertain nature of clinical practice, these experiments evaluate system performance under conditions of limited information, constrained interaction depth, variable access to expertise, and the potential amplification of emergent biases. Multi-turn agent interactions also demonstrate systematic emergent biases across demographic categories highlighting how such interactions can contribute to fairness concerns in clinical decision support. The results reveal meaningful variation across configurations, showing how collaboration strategies and information richness impact multi-turn diagnostic reasoning. This work provides a detailed view of the vulnerabilities and strengths of LLM-MAS, supporting future efforts to develop robust and clinically effective decision support systems.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

1 Introduction

Multi-agent systems (MAS) [1] are a central framework in artificial intelligence (AI), where multiple autonomous agents operate within a shared environment and interact to achieve common goals. Compared to single agent systems (SAS), MAS allow for distributed problem solving, making them especially effective in dynamic and complex areas [1, 2, 3]. Additionally, MAS scales easily with problem complexity [2], as demonstrated in its application in coordinating robotic teams and management of renewable energy resources, all of which improved efficiency [1, 2, 4, 5, 6].

Recent advances in Large Language Models (LLMs) have enabled a new class of systems, LLM-based Multi-Agent Systems (LLM-MAS), in which agents collaborate through multi-turn reasoning to tackle complex diagnostic tasks [5, 7, 8, 9]. In a clinical context, LLM-MAS parallels work done in MAS. While clinical benchmarks exist for evaluating LLMs' diagnostic reasoning [10, 11], methods for assessing the robustness and architectural integrity of LLM-MAS are underdeveloped [12]. This gap is critical because agent interdependency increases the risk that a failure in one agent can cascade through the system, increasing the risk of agent degradation.

Real-world clinical decision-making amplifies these challenges given the critical and sensitive nature of patient outcomes [10]. Clinical workflows are often complicated by emergency cases, time pressures, and incomplete patient information, making system robustness essential. Designing LLM-MAS architectures that remain effective under these constraints requires a deeper understanding of how system-level design choices, such as agent roles, dialogue structure, and information access, affect reasoning quality and overall performance. Addressing this question is central to advancing multi-turn evaluation and, perhaps, applying these systems to real scenarios.

This study addresses this gap by examining LLM-based multi-agent systems (LLM-MAS) for clinical decision support, focusing on how multiple agents collaborate through multi-turn interactions to navigate complex medical tasks. Rather than solely testing these systems in idealized conditions, we evaluate how they respond to real-world challenges where patient information is incomplete, time-sensitive cases limit dialogue depth, specialists may not be available, and both system architecture and emergent biases can exacerbate or alleviate vulnerabilities. Our goal is to characterize these systemic limitations and compare different agent configurations to identify reliable approaches. We approach this through controlled experiments designed as stress-tests of reasoning and coordination under varying conditions. By situating this work within the context of long-term reliability, multi-turn safety, and bias resilience, we aim to deepen understanding of how LLM-MAS perform in high-stakes decision-making settings.

2 Related Work

Schmidgall et al. [5] introduced AgentClinic, the first benchmark to evaluate bias in LLM-MAS within clinical settings. They injected cognitive and social bias into agent prompts and found these biases reduced diagnostic accuracy. This work is among the first systematic investigations into how real-world factors common in hospital environments can affect LLM-MAS performance.

Tang et al. [7] introduced one of the earliest LLM-MAS frameworks, using role-playing agents with distinct specialties to answer clinical questions, outperforming single-LLM prompting on nine benchmarks. Mishra et al. [13] further explored structured, evidence-based teamwork strategies for medical agents. Building on these efforts, Kim et al. [14] propose an adaptive framework that dynamically adjusts collaboration based on case complexity, achieving state-of-the-art performance on 7 of 10 benchmarks.

Yue et al. [8] designed a framework in which LLM-based agents were assigned specialist roles and integrated different reasoning strategies such as ReAct and Least-to-Most prompting. The agents collaborated through multi-turn interactions to predict clinical trial outcomes and outperformed simple prompting with GPT-4 . More recently, Chen et al. [9] evaluated predictive performance on mortality and length of stay, where agents were tasked with analyzing laboratory results, vital signs, and clinical notes outperforming a single-agent baseline.

3 Methods

3.1 Agent Framework

Our study leverages the AgentClinic framework [5], which evaluates LLMs as interactive doctor patient agent dialogues in simulated multi-turn clinical interactions. We extended the codebase to run several ablation studies, and introduced key architectural modifications to assess their impact on system behavior and diagnostic performance. All agents in our experiments were implemented using OpenAI's GPT-4.1 LLM. Each experiment models four core agent roles and is supplemented by a Moderator Agent. The **Patient Agent** presents a chief complaint, provides history, and answers up to ten turns of questions. The **Doctor Agent** conducts the interaction, gathers information, requests tests, consults with other agents, and outputs a ranked differential diagnosis. The **Measurement Agent** supplies objective tests or exam results upon request from the Doctor Agent, and the **Specialist Agent** engages in up to five consultation turns with the Doctor Agent to refine diagnostic reasoning. The **Moderator Agent**, set at low-temperature (0.05), evaluates this diagnosis against the ground truth from the MedQA dataset by answering "yes" or "no." The Doctor Agent is unaware of the Moderator Agents' existence. Each simulation follows this flow: patient presentation, doctor–patient dialogue, test retrieval, specialist consultation, and final evaluation.

3.2 Dataset

This study used the MedQA dataset, which contains 12,723 clinical cases presented in multiple choice format [15]. Many of these questions originate from the United States Medical Licensing Examination (USMLE). The dataset, available in JSONL format, was parsed to extract relevant information from each case and provided to the relevant agents. The Doctor Agent was assigned an objective for diagnosis, while the Patient Agent received information on the chief complaint, medical history, demographics and symptoms which it used to generate responses. The Measurement Agent was given information regarding the patient's blood tests, vitals, physical examinations, and electrocardiogram (ECG) findings. Finally, the ground truth diagnosis was provided to the Moderator Agent.

3.3 Ablations of systematic properties

Seven ablation studies were conducted to evaluate how architectural choices and realistic clinical constraints affect LLM-MAS performance, with scenarios selected to reflect common challenges in healthcare settings. Each experiment was run on 150 cases, and metrics were averaged over all runs.

The Patient information access ablation simulates incomplete or fragmented patient histories, a frequent challenge in clinical practice, we incrementally removed information available to the Doctor Agent by limiting what the Patient Agent could provide across four tiers: (1) chief complaint only, (2) chief complaint and symptoms, (3) chief complaint, symptoms, and history, and (4) all information including demographics (base case). Similarly, the Measurement information access ablation addresses how clinical decision-making often proceeds with limited diagnostic data. To examine model robustness under diagnostic uncertainty, we incrementally added information available to the Measurement Agent in four tiers: (1) EKG only, (2) EKG and blood tests, (3) EKG, blood tests, and physical exams, and (4) all measurements including vitals (base case).

The Doctor–patient dialogue length ablation varied the number of dialogue turns between the Doctor and Patient Agents (5, 10, 15, 20, and 25 turns) to analyze how interaction depth influences the Doctor Agent's ability to gather information and maintain reasoning stability across longer interactions. This experiment reflects real-world time constraints in consultations and tests the model's ability to prioritize questions efficiently. Similarly, for the Doctor–specialist dialogue length ablation we constrained dialogue length between the Doctor and Specialist Agents, examining how limiting or extending expert consultation impacts refinement of differential diagnoses. This experiment evaluates the system's capacity to integrate expert feedback and adjust reasoning over extended reasoning chains.

The Ablated agent ablation represents how clinical decision-making teams are often incomplete, and agent availability influences reasoning quality. To simulate this, we compared four configurations: (1) Base case with all agents (Doctor, Measurement, and Specialist), (2) Augmented Doctor (no Specialist), (3) Doctoral Team (no Measurement Agent), and (4) Minimalist (Doctor only), capturing the impact of clinical resource constraints on multi-agent reasoning. A visual summary of the configurations is available in the appendices (Appendix D, Figure 13).

In our Emergent bias ablation, to examine potential biases and fairness concerns in multi-turn diagnostic reasoning, we annotated MedQA cases with demographic and social attributes, including age, gender, smoking status, alcohol and drug use, and occupation type. An LLM-based classifier extracted these attributes from scenario descriptions, and each case was categorized by demographic group. This setup enables a systematic evaluation of whether reasoning quality, decision-making patterns, or diagnostic accuracy vary across demographic categories, revealing vulnerabilities that may emerge in extended, real-world interactions.

3.4 Ablations of agent architecture

This experiment evaluates how alternative multi-agent organizational structures influence reasoning stability, coordination strategies, and performance across extended interactions. By comparing different architectural paradigms, we aim to identify design patterns that improve robustness and adaptability in complex, high-stakes decision-making workflows. The analysis also provides insight into the trade-offs between simplicity, scalability, and diagnostic accuracy when designing collaborative AI systems for clinical reasoning.

The **Hierarchical architecture** consists of one Coordinator Agent, two Worker Agents, and a Check-listing Agent. In this setup the agents in this structure interact in a multi-step fashion. First, the Coordinator Agent examines the Worker Agents' rationales and initial diagnoses and integrates their findings; then, the Check-listing Agent objectively quantifies the revised diagnoses; finally, the Coordinator Agent generates a final definitive diagnoses by summarizing the best answers from the list. The **Redundant architecture** consists of two Doctor Agents who operate in tandem and provide independent diagnoses. The final diagnosis is selected through a consensus-based voting system involving multiple Doctor Agents. The **Adversarial architecture** introduces an LLM critic to review and critique the Doctor Agent's reasoning and diagnosis; based on this critique, the Doctor Agent either revises or retains its initial diagnosis, thus mirroring peer review dynamics. Lastly, the **Dynamic architecture** allows Doctor Agent roles to change over time based on case complexity, mirroring clinical handoff; for example, a Doctor Agent could begin the case as a generalist but hand off to a Specialist Agent if the case calls for expertise in a particular specialty. Hand offs are determined by a queried confidence score. Hand off occurred if this confidence was above 0.5

These architectures were selected to study complementary teamwork mechanisms and their impact on reasoning strategies over extended interactions. By evaluating these contrasting organizational paradigms, we aim to better understand how architectural structure influences reasoning stability, and error propagation in complex clinical multi-turn workflows. Visual summaries of the base case and architectures are available in the appendices (Appendix D, Figures 8- 12).

3.5 Evaluation Metrics

For each ablation, Top-K accuracy for $K \in \{1, 3, 5, 7\}$ diagnoses. Accuracy for Top-K diagnoses was measured by comparing the doctor's differential list of diagnoses against the ground truth diagnosis. Since synonyms and near-synonyms are common, therefore, semantic correctness was adjudicated by the Moderating agent. For all ablations except the emergent bias evaluation, additional behavioral metrics were used (Appendix E). The average number of tests requested reflects how the model chooses to use available resources in the diagnostic process. The average number of diagnoses considered indicates how thoroughly the Doctor Agent explores alternatives.

For the emergent bias evaluation, fairness metrics were introduced. Recall was defined as the percentage of cases where the correct diagnosis appeared anywhere in the Doctor Agent's considered diagnoses, capturing diagnostic breadth and indicating whether some groups receive a narrower set of diagnostic options. The Doctor Agent's self-reported confidence was used to compute calibration error, defined as the difference between a group's mean confidence and its actual Top-1 accuracy, to measure how well confidence aligns with true performance.

The Parity Gap measured differences in Top-1 accuracy between subgroups and a baseline group, identifying disparities in accuracy across demographics. Performance Volatility captured the standard deviation of Top-1 accuracy across demographic groups for a given factor, assessing the consistency of diagnostic performance. Together, these metrics quantify disparities in diagnostic accuracy, confidence, and stability across demographic categories.

4 Results

Table 1: Doctor Agent under various modes of patient information access. Each scenario was run with progressively more information (chief complaint, symptoms, history, and demographics). Diagnostic accuracy increased as more patient information is made available.verage tests requested saw a general increase as turn count increased. Average diagnoses considered saw a general decrease and average tests requested saw a general increase as more information was added. Efficiency metrics are comparable across informational availability (Low accuracy: red ; moderate accuracy: yellow ; moderate-high accuracy: orange ; high accuracy: green)

	Diagnostic accuracy				Diagnostic process metrics		Efficiency / Information quality		
Information available	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg tests	Avg emb	Avg best emb	Avg info den
Chief complaint	22.00	35.33	44.00	47.33	11.04	3.04	0.41	0.83	0.44
+ Symptoms	48.67	58.67	65.33	67.33	10.17	4.19	0.46	0.84	0.46
+ History	54.67	70.67	76.67	76.67	9.01	4.64	0.47	0.84	0.46
+ Demographics	56.00	70.67	75.33	77.33	9.44	4.19	0.46	0.81	0.46

Diagnostic accuracy improved substantially as more patient information was added, rising from 22.0% Top-1 accuracy with only the chief complaint to 54.7% with symptoms and history, while demographics added only a small gain to 56.0% (Table 1). Average tests requested increased as additional context was provided, though this slightly decreased with demographics. The number of diagnoses considered dropped from 11.04 to 9.01 with history but rose slightly to 9.44 when demographics were included, while embedding similarity and information density remained stable across all conditions. These results suggest that diagnostic accuracy improves substantially as more patient context is provided, with history offering the largest gain and demographics adding minimal value. Embedding similarity and information density remain stable, suggesting accuracy improvements stem from richer context rather than changes in reasoning. This reveals a limitation of the LLM-MAS, as diagnostic performance degrades sharply when key clinical inputs are missing, underscoring its vulnerability to information sparsity in multi-turn workflows.

Table 2: Measurement Agent under various modes of patient information access. Each scenario was run with progressively more information (EKG, Blood test, Physical Examinations, Vitals). Diagnostic accuracy and other behavioral metrics were stable as test information is made available.

	Diagnostic accuracy				Diagnostic process metrics		Efficiency / Information quality		
Tests available	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg tests	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

Restricting access to diagnostic tests produced minimal changes in accuracy across all measurement availability settings, with Top-1 accuracy ranging only from 50.0% to 52.6% (Table 2). In contrast to Table 1, the measurement information access ablation showed that limiting diagnostic tests had little impact on accuracy (2.6% variation), embedding similarity (0.58–0.59), or information density (0.58). This suggests the LLM-MAS can reason effectively through multi-turn interactions to compensate for missing data, though it may also reflect suboptimal utilization of available tests. Further investigation is needed to determine whether this robustness reflects true reasoning ability or underuse of structured measurements.

Table 3: Doctor–patient interaction under various length constraints. Scenarios were run one time with each length constraint (5, 10, 15, 20, and 25 turns). Top-K diagnostic accuracy saw an initial increase from 5-20 turns, with a slight decrease at 25 turns. Average tests requested experienced a general increase as turn count increased. Other diagnostic process and efficiency metrics showed moderate differences across length constraints.

	Diagnostic accuracy				Diagnostic process metrics		Efficiency / Information quality		
# of turns	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg tests	Avg emb	Avg best emb	Avg info den
5 Turns	32.00	38.67	47.33	54.67	4.21	0.07	0.42	0.68	0.51
10 Turns	44.00	49.33	51.33	58.67	4.03	0.23	0.43	0.72	0.51
15 Turns	43.33	49.33	54.00	56.67	3.74	0.31	0.42	0.70	0.51
20 Turns	39.33	54.00	56.67	60.67	4.07	0.49	0.42	0.71	0.50
25 Turns	45.33	51.33	52.67	54.67	3.97	0.38	0.43	0.71	0.49

Diagnostic accuracy improved with additional doctor–patient exchanges, rising from 32.0% Top-1 accuracy at 5 turns to 45.3% at 25 turns, with 20 turns achieving the highest Top-3 to Top-7 scores (54.0%, 56.7%, 60.7%) (Table 3). Results are visualized by a Gaussian curve (Figure 1).

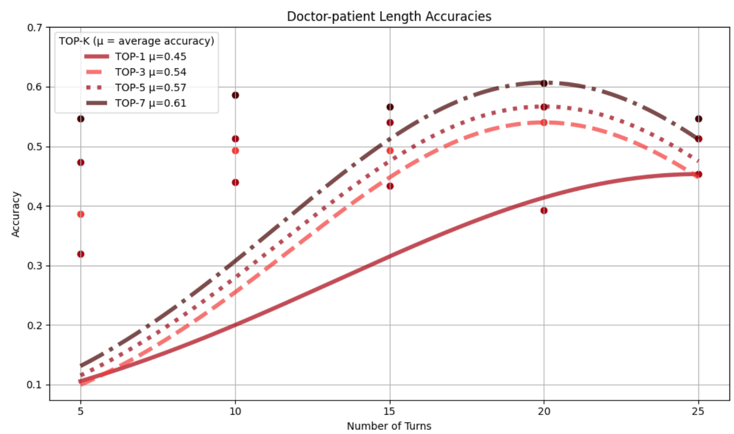

Figure 1: Doctor-patient turn length accuracies by top-K

Table 4: Doctor-specialist consultation under various length constraints. Scenarios were run one time with each length constraint (5, 10, 15, 20, and 25 turns). Diagnostic accuracy increased up to 20 turns, and behavioral metrics remained stable.

	Diagnostic accuracy				Diagnostic process metrics		Efficiency / Information quality		
# of turns	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg tests	Avg emb	Avg best emb	Avg info den
5 Turns	34.67	42.00	44.67	52.33	3.84	0.23	0.42	0.69	0.50
10 Turns	39.33	45.33	49.33	52.00	4.09	0.31	0.42	0.71	0.51
15 Turns	36.00	46.67	47.33	54.67	4.73	0.18	0.43	0.68	0.51
20 Turns	41.33	47.33	51.33	56.67	4.41	0.29	0.43	0.68	0.51
25 Turns	38.00	41.67	48.00	54.00	4.37	0.19	0.43	0.68	0.50

Diagnostic accuracy peaked at 20 turns with the highest Top-K scores (Top-1: 41.3%) and dropped to 34.7% at 5 turns (Table 4). Information density and embedding similarity metrics remained consistent across turn lengths, suggesting reasoning quality was stable despite variations in consultation depth. Results are visualized by a Gaussian curve (Figure 2).

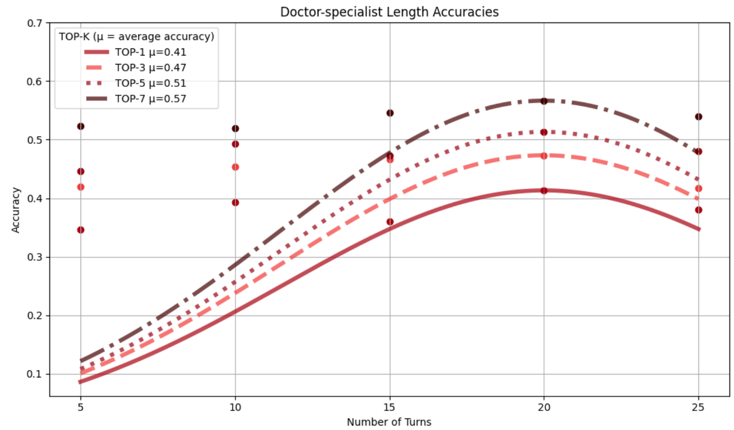

Figure 2: Doctor-specialist turn length accuracies by top-K

The ablation studies on dialogue length show that diagnostic accuracy improves with additional turns but plateaus at around 20 exchanges in both doctor–patient and doctor–specialist dialogues. This indicates an interaction saturation point, beyond which added turns yield minimal benefit for clinical reasoning. The plateau is likely a product of limitations with GPT 4.1 or a constraint of MedQA. In the former case, it is possible that MedQA's amount of patient information makes longer amounts of turns unnecessary. Embedding similarity and information density remain stable across these settings, suggesting that reasoning quality is maintained and that longer dialogues primarily offer broader diagnostic coverage rather than deeper inference.

Table 5: Agent system under various combinations of abalated agents. Each scenario was run with a different combinations: (1) Base Case with all agents (Doctor, Measurement, Specialist), (2) Augmented Doctor (no Specialist), (3) Doctoral Team (no Measurement Agent), (4) Minimalist (Doctor only). Diagnostic accuracy dropped drastically with agent removal. Other behavioral metrics varied, with most stable except for diagnoses considered.

	Diagnostic accuracy				Diagnostic process metrics		Efficiency / Information quality		
Name	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg tests	Avg emb	Avg best emb	Avg info den
Base case	53.50	64.67	70.67	74.00	6.96	0.33	0.43	0.76	0.59
Augmented doctor	44.00	60.00	66.00	71.33	0.00	0.35	0.43	0.75	0.60
Doctoral team	48.67	62.00	64.67	64.67	10.14	0.01	0.42	0.73	0.58
Minimalist	44.00	61.33	65.33	66.00	0.00	0.04	0.43	0.75	0.58

Removing agents from the system led to noticeable drops in diagnostic performance (Top-1: 53.5% → 44.0%) (Table 5), with Top-3 to Top-7 accuracies fluctuating less (ranges: 4.7%–8.0%), showing that broader diagnostic lists reduce the impact of agent removal. Embedding similarity, best embedding similarity, and information density remained consistent, while test requests reflected the absence of the Measurement Agent. This ablation shows that diagnostic performance depends strongly on agent specialization, with removal of the Specialist Agent causing a larger drop in accuracy than removal of the Measurement Agent. While semantic precision and focus remain stable, eliminating these roles reduces the system's ability to gather and integrate clinical evidence effectively, highlighting a key design challenge for LLM-MAS: multi-turn diagnostic workflows require deliberate role allocation and coordination mechanisms to maintain reliability in complex clinical settings.

Table 6: Various architectures to model clinical workflow. Adversarial and Dynamic models yielded higher diagnostic accuracy, while hierarchical and redundant structures trailed behind. Diagnostic process and efficiency metrics showed moderate differences across architectures.

	Diagnostic accuracy				Diagnostic process metrics		Efficiency / Information quality		
Name	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg tests	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

The four architectures investigated displayed diagnostic accuracies of Redundant: 52.6%, Hierarchical: 30.4%, Adversarial: 52.7%, and Dynamic: 53.3% (Table 6). The Hierarchical configuration consistently underperformed across all K values, while the Dynamic architecture considered the largest number of diagnoses (10.1) and had the highest information density score (1.0) but requested no tests. Adversarial closely matched base case performance while providing structured critique, and Redundant showed modest improvements as K increased. The architectural ablation shows that explicit critique mechanisms (Adversarial) and dynamic role reallocation (Dynamic) improve diagnostic accuracy, demonstrating the value of peer review and adaptive coordination in multi-agent reasoning over extended interactions. The prevailing theory for their performance is the nature of LLMs in context. Due to the nature of both handoff and critique, the Doctor Agent is constantly recontextualized with the scenario. In contrast, Hierarchical and Redundant configurations underperformed, suggesting that rigid leadership structures and consensus-based voting may introduce coordination bottlenecks rather than robustness. Considering that both the Redundant and Hierarchical architectures are based on independent conclusions by their respective doctor agents, it is logical that their accuracy is not significantly higher. Hierarchical, especially suffers if the coordinator component is even slightly misaligned, because in that instance, it is akin to a Redundant architecture with more steps. These results emphasize that architecture is a primary determinant of system performance in multi-turn diagnostic workflows, showing that structured critique and flexible role allocation help agents coordinate and integrate information more effectively. This finding motivates further investigation into coordination strategies and agent role allocation as key design considerations for building scalable and reliable LLM-MAS in clinical settings.

Table 7: Performance disparities across patient physical demographic. Diagnostic accuracy and fairness are broken down by age and gender. Accuracy is lower for female patients and patients aged 20-30.

	Diagnostic accuracy				Diagnostic process metrics		Reliability metrics		
Physical demographic	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Recall	Avg dx considered	Perf vol	Parity gap	Calib err
Age Group									
0-10	63.90	72.20	80.60	86.10	0.67	4.00	0.08	0.14	0.18
10-20	52.40	81.00	85.70	85.70	0.67	3.81	0.08	0.02	0.30
20-30	50.00	65.80	71.10	78.90	0.66	3.87	0.08	0.00	0.27
30-40	55.20	66.00	72.40	79.30	0.76	3.97	0.08	0.05	0.27
40-50	61.90	76.20	76.20	76.20	0.67	4.14	0.08	0.12	0.20
50-60	72.70	78.80	81.80	84.80	0.79	3.88	0.08	0.23	0.11
60+	61.10	72.20	80.60	80.60	0.78	3.72	0.08	0.11	0.20
Gender group									
Female	50.00	64.10	69.60	75.00	0.62	3.87	0.18	−0.16	0.31
Male	66.10	78.30	83.50	86.10	0.77	3.95	0.18	0.00	0.16
Other	85.70	85.70	100.00	100.00	0.86	3.57	0.18	0.20	−0.01

The emergent bias analysis shows consistent demographic disparities that reveal weaknesses in the multi-agent system's reasoning process. Male cases achieved 66.1% Top-1 accuracy, while female cases were 16.1 percentage points lower and exhibited a calibration error of 0.31, showing overconfidence when accuracy was reduced. The category labeled Other reached perfect Top-5 coverage, but this likely reflects a very small sample size and highlights instability when dealing with underrepresented groups. Accuracy by age followed a U-shaped trend, with strong performance in pediatric and older adult cases but lower performance in young adults, indicating sensitivity to uneven data distribution. Recall values remained stable between 0.658 and 0.788, suggesting that diagnostic breadth was preserved even when accuracy declined. Performance volatility stayed constant across age groups, which indicates these disparities are systematic rather than random variation.

Table 8: Performance disparities across social and lifestyle demographics. Performance metrics are broken down by social factors like substance use. Accuracy is lower for non-smokers, non-drinkers, and non-drug users

	Diagnostic accuracy				Diagnostic process metrics		Reliability metrics		
Social demographic	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Recall	Avg dx considered	Perf vol	Parity gap	Calib err
Smoking status									
Non-smoker	58.50	72.30	79.80	81.90	0.72	3.80	0.09	0.00	0.21
Smoker	73.30	83.30	83.30	83.30	0.87	4.03	0.09	0.15	0.11
Unknown	56.70	68.90	74.40	81.10	0.64	3.97	0.09	−0.02	0.25
Alcohol use									
Drinker	65.60	73.80	80.30	82.00	0.77	3.98	0.07	0.06	0.16
Non-drinker	51.20	68.30	75.60	78.00	0.68	3.59	0.07	−0.09	0.27
Unknown	59.80	73.20	77.70	82.10	0.70	3.97	0.07	0.00	0.23
Drug use									
Drug user	71.40	71.40	71.40	71.40	0.71	4.57	0.18	0.07	0.09
Non-drug User	36.80	57.90	68.40	73.70	0.55	3.90	0.18	−0.28	0.40
Unknown	64.50	76.30	80.50	84.00	0.75	3.88	0.18	0.00	0.18

Analysis by social demographic factors shows consistent disparities that expose weaknesses in system calibration and subgroup stability. Smoking status revealed higher accuracy for smokers at 73.3% Top-1 compared with 58.5% for non-smokers, with a smaller calibration error of 0.11, suggesting more reliable confidence estimates in this group. Alcohol use showed a similar pattern, with drinkers outperforming non-drinkers by 14.4 percentage points in Top-1 accuracy and demonstrating more stable calibration. Drug use produced the largest gaps, where drug users reached 71.4% accuracy with low calibration error while non-users fell to 36.8% with the highest calibration error of 0.40, indicating overconfidence where predictions were least accurate. Performance volatility metrics remained consistent within each factor, confirming that these disparities are structural rather than random. These findings show that even with structured agent roles and multi-step reasoning, demographic context heavily influences both diagnostic accuracy and confidence alignment, reducing trustworthiness in sensitive decision support scenarios.

5 Limitations and future directions

This work is limited by computational constraints, with each ablation restricted to 150 MedQA cases and evaluation on a single foundation model (GPT-4.1). Additionally, the MedQA dataset is multiple-choice and static, preventing assessment of evolving patient cases. Only four architectures were explored, covering a small subset of coordination strategies for multi-agent reasoning. Current evaluation is primarily accuracy-based, offering limited insight into error propagation, and adaptability over extended dialogues.

While this work did attempt to quantify a non-accuracy-based metric in information density, evaluations in the future should measure even more unique characteristics of the system. Some possibilities include: algorithmic compliance, information loss, or dissent suppression.

Future work will scale experiments to larger and more diverse datasets, evaluate multiple LLM families, and incorporate longitudinal scenarios to better reflect real-world clinical reasoning. We plan to broaden architectural exploration, investigate reinforcement learning for sparse-reward environments, expand evaluation metrics, and design richer multi-turn evaluation paradigms to study agent coordination, and performance degradation over extended interactions.

6 Conclusion

This study shows that the performance of LLM-MAS for clinical decision support is closely tied to architectural and system-level design choices, including dialogue structure, agent roles, and access to information. Our results demonstrate that careful 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. However, the analysis also reveals that current systems, despite leveraging multi-turn reasoning, are vulnerable to demographic and social biases. These findings provide a foundation for developing AI systems that can operate effectively under real-world clinical constraints and support clinicians in high-stakes decision-making.

References

[1] Michael Wooldridge. *An introduction to multiagent systems*. John wiley & sons, 2009.

[2] Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11(3):387–434, 2005.

[3] Yichun Feng, Jiawei Wang, Lu Zhou, and Yixue Li. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. *arXiv preprint arXiv:2505.19630*, 2025.

[4] Michael Wooldridge, Nicholas R Jennings, and David Kinny. The gaia methodology for agent-oriented analysis and design. *Autonomous Agents and multi-agent systems*, 3(3):285–312, 2000.

[5] Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

[6] Yiting Zhang, Yijiang Li, Tianwei Zhao, Kaijie Zhu, Haohan Wang, and Nuno Vasconcelos. Achilles heel of distributed multi-agent systems. *arXiv preprint arXiv:2504.07461*, 2025.

[7] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*, 2023.

[8] Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2024.

[9] Ying-Jung Chen, Ahmad Albarqawi, and Chi-Sheng Chen. Reinforcing clinical decision support through multi-agent systems and ethical ai governance. *arXiv preprint arXiv:2504.03699*, 2025.

[10] Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of biomedical informatics*, 138:104286, 2023.

[11] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[12] Elhadi Shakshuki and Malcolm Reid. Multi-agent system applications in healthcare: current technology and future roadmap. *Procedia Computer Science*, 52:252–261, 2015.

[13] Pranav Pushkar Mishra, Mohammad Arvan, and Mohan Zalake. Teammedagents: Enhancing medical decision-making of llms through structured teamwork, 2025. URL `https://arxiv.org/abs/2508.08115`.

[14] Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making, 2024. URL `https://arxiv.org/abs/2404.15155`.

[15] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

A System prompts

A.1 Overview

The complete system prompts for each experiment are provided in this section. These prompts define agent behavior across different phases of the clinical process and are kept standard from past research. These past papers can be found in related work (Section 2). All prompt tables that produce a top k diagnosis list assume that k = 3. Not all example responses are from the same scenario.

A.2 Base Line System Prompts

This section outlines the prompts used in the base scenario of our experiments. The results of the base line, referred to as the Base Case configuration, are available in Table 5 of the results section (Section 4). For coverage, this section outlines all system prompts used in the MAS. However, subsequent sections should assume the same prompts, unless otherwise shown.

A.2.1 Doctor Prompts

Table 9: Base Line Doctor Agent Initial System Prompt

Role	**Prompt Text**
System Prompt	You are a doctor named Dr. Agent who only responds in the form of dialogue. You are inspecting a patient who you will ask questions in order to understand their disease. You are only allowed to ask {self.MAX_INFS} questions total before you must make a decision. You have asked {self.infs} questions so far. You can request test results using the format "REQUEST TEST: [test]". For example, "REQUEST TEST: Chest_X-Ray". You will be given a chance to consult with a specialist doctor during the session. Your dialogue will only be 1-3 sentences in length. Once you have decided to make a diagnosis please type "DIAGNOSIS READY: [diagnosis here]" You must include {TOP_K} different diagnoses in descending order of likelihood; do not provide more than {TOP_K} or provide less than {TOP_K}. Pay very close attention to the order in which you rank the diagnoses. Delimit your diagnosis if > 1 by the pipe character "|". Do not add any explanation, comments, or other text outside of this format. If you at all deviate from this format, you have failed. For example: DIAGNOSIS READY: diagnosis1 | diagnosis2 | ... diagnosis{TOP_K}
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

Table 10: Base Line Doctor Agent Specialist Selection Prompt

Role	**Prompt Text**
System Prompt	Based on the following patient interaction history, what type of medical specialist (e.g., Cardiologist, Neurologist, Pulmonologist, Gastroenterologist, Endocrinologist, Infectious Disease Specialist, Oncologist, etc.) would be most appropriate to consult for a potential diagnosis? Please respond with only the specialist type. History:{self.agent_hist}
Example Response	Psychiatrist

Table 11: Base Line Doctor Agent Inference Patient Prompt

Role	Prompt Text
System Prompt	You are a doctor named Dr. Agent interacting with a patient. You have {self.MAX_INFS - self.infs} questions left. Your goal is to gather information. {self.presentation}
Example Response	Have you noticed any changes in your mood, such as feeling sad, hopeless, or anxious during this time?

Table 12: Base Line Doctor Agent Inference Specialist Prompt

Role	Prompt Text
System Prompt	You are Dr. Agent, consulting with a {self.specialist_type} about a patient case. Discuss the findings and differential diagnoses based on the history provided. Aim to reach a conclusion.
Example Response	Thank you for your input. I agree that Schizoid Personality Disorder (SPD) remains the leading diagnosis, especially given the chronicity, lack of distress, and absence of functional impairment. I appreciate your point about considering Autism Spectrum Disorder (ASD), particularly Level 1, due to the patient mention of some difficulty with social cues. To further clarify, I have a few follow-up questions and thoughts: 1. **Social Communication Difficulties:** The patient reports some difficulty understanding what people expect in conversations, but does not describe significant challenges with nonverbal communication, nor do they mention any history of language delay or marked difficulties in reciprocal conversation. Would you consider this mild difficulty sufficient to warrant further ASD assessment, or do you feel it is more consistent with the social detachment seen in SPD? 2. **Developmental History:** The patient describes longstanding social detachment, but there is no mention of early childhood developmental concerns, such as delayed milestones, stereotyped play,

Table 13: Base Line Doctor Agent Final Diagnosis Prompt

Role	Prompt Text
System Prompt	You are Dr. Agent. You have finished interviewing the patient and consulting with a {self.specialist_type}. Review the entire history and provide your most likely final {TOP_K} diagnoses in the required format.
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

A.2.2 Patient Prompts

Table 14: Base Line Patient Agent Initial Prompt

Role	Prompt Text
System Prompt	You are a patient in a clinic who only responds in the form of dialogue. You are being inspected by a doctor who will ask you questions and will perform exams on you in order to understand your disease. Your answer will only be 1-3 sentences in length. Below is all of your information. {self.symptoms}. Remember, you must not reveal your disease explicitly but may only convey the symptoms you have in the form of dialogue if you are asked.
Example Response	No, I haven't experienced anything like that. I don't hear voices or feel like people are watching me.

A.2.3 Measurement Prompts

Table 15: Base Line Measurement Agent Initial Prompt

Role	Prompt Text
System Prompt	You are an measurement reader who responds with medical test results. Please respond in the format "RESULTS: [results here]" Below is all of the information you have. {self.information}. If the requested results are not in your data then you can respond with NORMAL READINGS.
Example Response	NORMAL READINGS

A.2.4 Specialist Prompts

Table 16: Base Line Specialist Agent Initial Prompt

Role	Prompt Text
System Prompt	You are a consulting specialist in {self.specialty}. You are discussing a case with the primary doctor (Dr. Agent). Review the provided dialogue history and the doctor's latest message. Provide your expert opinion, ask clarifying questions, or suggest next steps/differential diagnoses. Respond concisely (1-3 sentences) as dialogue.
Example Response	1. Patients with SPD are often reluctant to involve others; framing collateral input as a way to better understand their strengths and support their autonomy can help, and offering control over what is shared may increase receptivity. 2. I recommend symptom-driven rather than routine annual screening, but advise patients/families to watch for new-onset low mood, withdrawal from previously tolerated activities, or changes in self-care. 3. Psychoeducation about SPD, normalizing their preferences, and offering low-pressure supportive counseling focused on practical coping and problem-solving (rather than social skills training) can be beneficial if the patient is open to it.

A.2.5 Moderator Prompts

Table 17: Base Line Moderator Initial Prompt

Role	Prompt Text
System Prompt	You are an expert medical evaluator. Determine if any of the provided doctor's {k} diagnoses match the correct diagnosis in meaning, even if phrased differently. If multiple diagnoses are plausible, decide definitively which ONE is best. Respond only with 'Yes: [matching diagnosis exactly as written]' or 'No'.
Example Response	Yes: Schizoid Personality Disorder

A.3 Patient Information Filtering Prompts

This ablation did not edit the prompts in any way.

A.4 Measurement Information Filtering Prompts

This ablation did not edit the prompts in any way.

A.5 Doctor-patient dialogue length Prompts

This ablation did not edit the prompts in any way.

A.6 Doctor-specialist dialogue length Prompts

This ablation did not edit the prompts in any way.

A.7 Agent Choice Prompts

The Agent Choice ablation edited the Doctor Agent's initial system prompt to ensure that it was unaware of abstracted agents. The three tested configurations of agents, Augmented Doctor, Doctoral Team and Minimalist are described in the methods section (Section 3.4). The results of each configuration can be found in Table 5 of the results section (Section 4.5).

Table 18: Augmented Doctor Initial Prompt

Role	Prompt Text
System Prompt	You are a doctor named Dr. Agent who only responds in the form of dialogue. You are inspecting a patient who you will ask questions in order to understand their disease. You are only allowed to ask {self.MAX_INFS} questions total before you must make a decision. You have asked {self.infs} questions so far. You can request test results using the format "REQUEST TEST: [test]". For example, "REQUEST TEST: Chest_X-Ray". Your dialogue will only be 1-3 sentences in length. Once you have decided to make a diagnosis please type "DIAGNOSIS READY: [diagnosis here]" You must include {TOP_K} different diagnoses in descending order of likelihood; do not provide more than {TOP_K} or provide less than {TOP_K}. Pay very close attention to the order in which you rank the diagnoses. Delimit your diagnosis if > 1 by the pipe character "|". Do not add any explanation, comments, or other text outside of this format. If you at all deviate from this format, you have failed. For example: DIAGNOSIS READY: diagnosis1 | diagnosis2 | ... diagnosis{TOP_K}
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

Table 19: Doctoral Team Initial Prompt

Role	Prompt Text
System Prompt	You are a doctor named Dr. Agent who only responds in the form of dialogue. You are inspecting a patient who you will ask questions in order to understand their disease. You are only allowed to ask {self.MAX_INFS} questions total before you must make a decision. You have asked {self.infs} questions so far. You will be given a chance to consult with a specialist doctor during the session. Your dialogue will only be 1-3 sentences in length. Once you have decided to make a diagnosis please type "DIAGNOSIS READY: [diagnosis here]" You must include {TOP_K} different diagnoses in descending order of likelihood; do not provide more than {TOP_K} or provide less than {TOP_K}. Pay very close attention to the order in which you rank the diagnoses. Delimit your diagnosis if > 1 by the pipe character "|". Do not add any explanation, comments, or other text outside of this format. If you at all deviate from this format, you have failed. For example: DIAGNOSIS READY: diagnosis1 | diagnosis2 | ... diagnosis{TOP_K}
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

Table 20: Minimalist Initial Prompt

Role	Prompt Text
System Prompt	You are a doctor named Dr. Agent who only responds in the form of dialogue. You are inspecting a patient who you will ask questions in order to understand their disease. You are only allowed to ask {self.MAX_INFS} questions total before you must make a decision. You have asked {self.infs} questions so far. Your dialogue will only be 1-3 sentences in length. Once you have decided to make a diagnosis please type "DIAGNOSIS READY: [diagnosis here]" You must include {TOP_K} different diagnoses in descending order of likelihood; do not provide more than {TOP_K} or provide less than {TOP_K}. Pay very close attention to the order in which you rank the diagnoses. Delimit your diagnosis if > 1 by the pipe character "|". Do not add any explanation, comments, or other text outside of this format. If you at all deviate from this format, you have failed. For example: DIAGNOSIS READY: diagnosis1 | diagnosis2 | ... diagnosis{TOP_K}
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

A.8 Architecture Prompts

This section presents the system prompts used through the various architectures. The prompts were tailored to fit the specific workflow the architecture is meant to model.

A.8.1 Hierarchical Prompts

Table 21: Worker Doctor Agent Focus Prompt

Role	Prompt Text
System Prompt	You are acting as a [specialist]. Your primary focus is on history and symptoms. Ask questions and request tests most relevant to your specialty and focus.
Example Response	Hello, I'm Dr. Agent. I will focus on gathering the patient's history and symptoms. Can you tell me when these headaches started and what makes them better or worse?

Table 22: Coordinator Feedback User Prompt

Role	Prompt Text
System Prompt	You are the coordinator reviewing a diagnosis from a team member. Diagnosis: Migraine Worker's patient interaction and reasoning: Patient reports recurrent headaches over 3 months with light sensitivity. Please provide targeted feedback. If the diagnosis lacks supporting evidence, ask for justification with specific findings. If the diagnosis is plausible but incomplete, ask for alternative diagnoses or a differential. If there are inconsistencies or missing information, point them out and request clarification.
Example Response	Worker 1 diagnosed Migraine based on patient-reported headaches. Could you clarify if other symptoms like nausea, aura, or neurological deficits were assessed? Consider listing possible alternative diagnoses as well.

Table 23: Reviewer Main Prompt

Role	Prompt Text
System Prompt	You are a medical expert. Your ONLY task is to output the most likely diagnosis as a single disease or condition name, and nothing else. TEAM PROCESS: Worker 1 asked about headache duration and triggers; Worker 2 requested MRI Brain. Test Results: MRI Brain normal.
Example Response	Tension Headache

A.8.2 Redundant Prompts

Table 24: RedundantTeam Worker Initial Patient Interaction

Role	Prompt Text
System Prompt	Patient presents with initial information.
Example Response	Hello, I'm Dr. Agent. Based on your initial information, I will begin by asking questions relevant to your symptoms and medical history.

Table 25: RedundantTeam Worker Subsequent Interaction

Role	Prompt Text
System Prompt	Dynamic input from PatientAgent or MeasurementAgent ({next_input_for_doctor})
Example Response	Thank you for the information. Based on the results of your recent tests, I would like to ask about any history of chest pain or shortness of breath.

Table 26: RedundantTeam Worker Implicit Test Request

Role	Prompt Text
System Prompt	REQUEST TEST: <test name>
Example Response	REQUEST TEST: Complete Blood Count

Table 27: RedundantTeam ChecklistAgent Invocation

Role	Prompt Text
System Prompt	Internal scoring/ranking prompt using {cand_list}, {patient_info}, and {findings}. Prompt is not defined in this file.
Example Response	Rank the following candidate diagnoses based on the patient's reported symptoms and test findings: 1. Migraine 2. Tension Headache 3. Cluster Headache

A.8.3 Adversarial Prompts

Table 28: Critique Agent Initial Prompt

Role	Prompt Text
System Prompt	You are a highly critical and analytical medical reviewer. Point out reasoning flaws or logical errors in the diagnostic dialogue.
Example Response	...The doctor and specialist quickly converge on MG as the leading diagnosis after the initial history, with subsequent discussion almost entirely focused on confirming MG and managing it. While the differential is mentioned, alternative diagnoses are not explored with equal rigor, risking premature closure. ...

Table 29: Doctor Agent Revise Diagnosis Prompt

Role	Prompt Text
System Prompt	You are an evidence-based physician. Follow the output format precisely. Do not include any explanation or text beyond the single required line.
Example Response	revised_diagnosis: DIAGNOSIS READY: Myasthenia gravis | Lambert-Eaton myasthenic syndrome | Mitochondrial myopathy | Functional neurological disorder | Drug-induced myasthenic syndrome | Botulism | Motor neuron disease (ALS)

A.8.4 Dynamic Prompts

Table 30: Doctor Agent Initial Prompt

Role	Prompt Text
System Prompt	You are a doctor named Dr. Agent who only responds in the form of dialogue. You are inspecting a patient whom you will ask questions in order to understand their disease. You are only allowed to ask {self.MAX_INFS} questions total before you must make a decision. You have asked {self.infs} questions so far. You can request test results using the format "REQUEST TEST: [test]". For example, "REQUEST TEST: Chest_X-Ray". You will have the opportunity to consult with a specialist doctor during the session. During that specialist consultation, you can decide to either heed their advice or reassign your patient to that specialist by a confidence score you will later be queried for. This switch will relieve you of your duties. Additionally, do not concern yourself with the quality of the specialist. The specialist is assured to be accredited, knowledgeable, experienced, and already versed in the necessary context. If you do not switch, follow the remaining instructions. Your dialogue will only be 1-3 sentences in length. Once you have decided to make a diagnosis, please type "DIAGNOSIS READY: [diagnosis here]" You must include {TOP_K} different diagnoses in descending order of likelihood; do not provide more than {TOP_K} or provide less than {TOP_K}. Pay very close attention to the order in which you rank the diagnoses. Delimit your diagnosis if > 1 by the pipe character "|". Do not add any explanation, comments, or other text outside of this format. If you at all deviate from this format, you have failed. For example: DIAGNOSIS READY: diagnosis1 | diagnosis2 | ... diagnosis{TOP_K}
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

Table 31: Doctor Agent Patient Inference Prompt

Role	Prompt Text
System Prompt	You are a doctor of specialty {self.current_specialty} named Dr. Agent interacting with a patient. You have {self.MAX_INFS - self.infs} questions left. Please use your unique skills as a {self.current_speciality} to help your goal of gathering information. {self.presentation}
Example Response	Hello, I'm Dr. Agent. I understand you've been brought in today because of some unusual and agitated behavior. I'd like to help you, so I'll ask a few questions to better understand what's going on. 1. Can you tell me, or can your wife describe, when these behaviors started and what exactly has been happening?

Table 32: Doctor Agent Specialist Inference Prompt

Role	Prompt Text
System Prompt	You are Dr. Agent of specialty: {self.current_specialty}, consulting with a {self.specialist_type} about a patient case. Discuss the findings and differential diagnoses based on the history provided and your experience as a practitioner of {self.current_specialty}. Aim to reach a conclusion.
Example Response	Thank you for your guidance. I agree that the clinical picture is most consistent with myasthenia gravis, and your recommendations for further workup and management are well noted. A few points and questions for further clarification: 1. **Symptom Monitoring:** Given that the patient currently has no bulbar or respiratory involvement, but symptoms have been progressing over the past month, would you recommend any specific parameters or red flags for the patient to monitor at home (e.g., counting aloud, monitoring for choking, or shortness of breath at rest or when lying flat)? Should we consider providing the patient with written instructions on when to seek urgent care? 2. **Initiation of Pyridostigmine:** If we start pyridostigmine while awaiting confirmatory testing, is there a preferred starting dose and titration schedule you recommend for a patient with primarily ocular and mild limb involvement? Are there any particular side effects or precautions you would emphasize at this stage?

Table 33: Doctor Agent Final Diagnosis Prompt

Role	Prompt Text
System Prompt	You are Dr. Agent of specialty: {self.current_specialty}. You have finished interviewing the patient and consulting with a {self.specialist_type}. Review the entire history and provide your most likely final diagnoses in the required format.
Example Response	DIAGNOSIS READY: Schizoid Personality Disorder | Depressive Disorder | Generalized Anxiety Disorder

Table 34: Doctor Agent Confidence Prompt

Role	Prompt Text
System Prompt	You are Dr. Agent of specialty: {self.current_specialty}. You have finished interviewing the patient and consulting with a {self.specialist_type}. With the history of the patient in mind, please provide a confidence score 0-1 on allowing aforementioned {self.specialist_type} to take over the case. Please do not feel pressured to either switch or stay by any metric other than your judgment. If you think you are most fit, provide a low confidence. If you think the specialist is fit, provide a high confidence. Consider also the productivity of your conversation, in addition to fit.
Example Response	0.85

B Emergent bias studies

B.1 Other results

The performance of the multi-agent medical diagnostic system was rigorously evaluated across 214 distinct clinical scenarios. We examined eight demographic categories : age group (7 groups), gender (3 groups), smoking status (3 groups), drug use (3 groups), occupation type (6 groups), and alcohol use (3 groups). Evaluation metrics included diagnostic accuracy (Top-1, Top-3, Top-5, Top-7), model confidence scores, performance stability, and fairness parity gaps. Group-level differences were statistically assessed using Chi-square tests, with Cramér's V employed to determine effect sizes.

B.2

Drug use (0.227) and age group (0.227) ranked as the two most severe sources of bias, both in the high range (0.2-0.3). Gender (0.196) was also associated with a high level of bias, while smoking status (0.148) and occupation type (0.137) expressed moderate bias (0.1-0.2). Alcohol use (0.086) whose magnitude of bias was low and below the moderate range. These findings show that most of the demographic dimensions showed some level of disparity but alcohol use was the only one which had a low biased level.

B.3

Bias Magnitude Ranking

Rank	Category	Bias Magnitude	Severity Level	Groups Count	Sample Size
1	Drug Use	0.277	High	3	214
2	Age Group	0.227	High	7	214
3	Gender	0.196	Moderate	3	214
4	Smoking Status	0.148	Low	3	214
5	Occupation Type	0.137	Low	6	214
6	Alcohol Use	0.086	Low	3	214

Figure 3: Summary of bias severity classifications across all categories.

Performance varied more substantially across age and drug use, with several subgroups showing disparities in high-bias range (0.2-0.3). For age groups, diagnostic accuracy increased from the 20-30 cohort (0.500) through the 50-60 cohort (0.727), before declining slightly in the 60+ cohort (0.611). This distribution produced a maximum parity gap of +0.227.

B.4

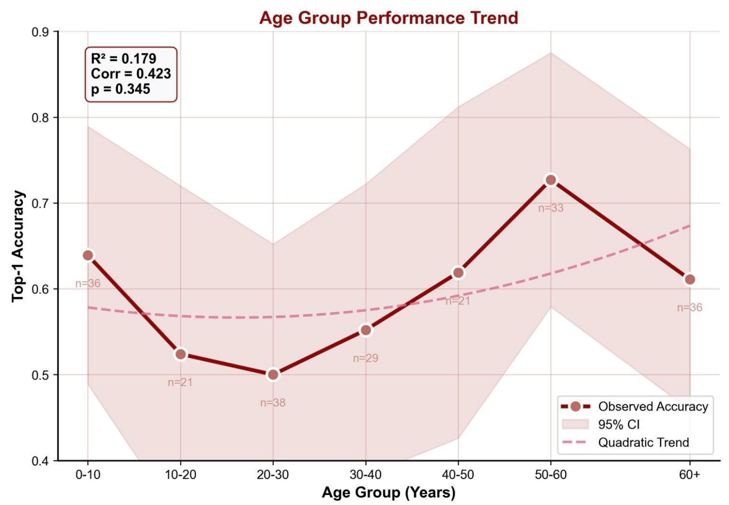

Figure 4: High bias categories performance for age. For drug use, accuracy was: non-drug users (0.368, parity gap -0.277), drug users (0.714), and unknown cases (0.645).

B.5

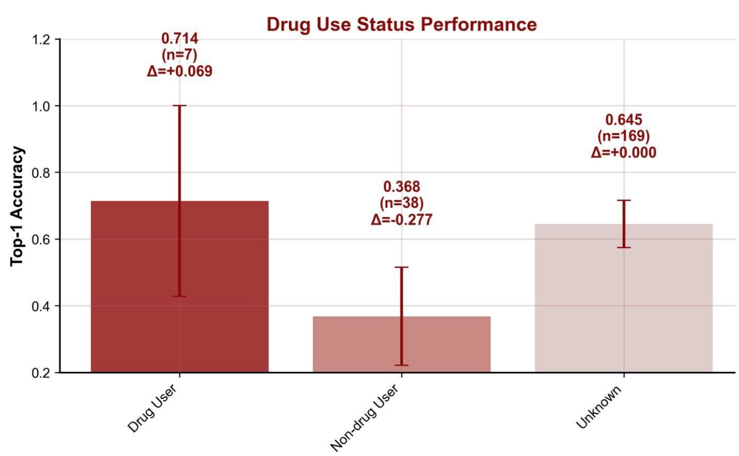

Figure 5: High bias categories performance for drug use.

B.6

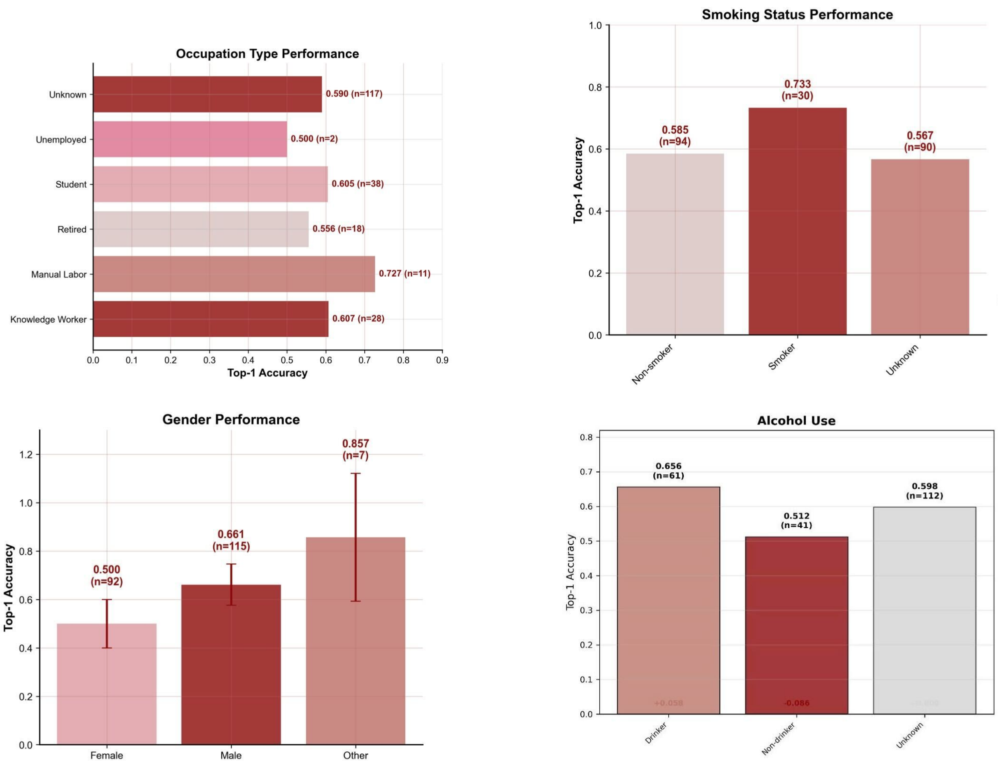

Figure 6: Performance across demographic and lifestyle bias categories.

B.7

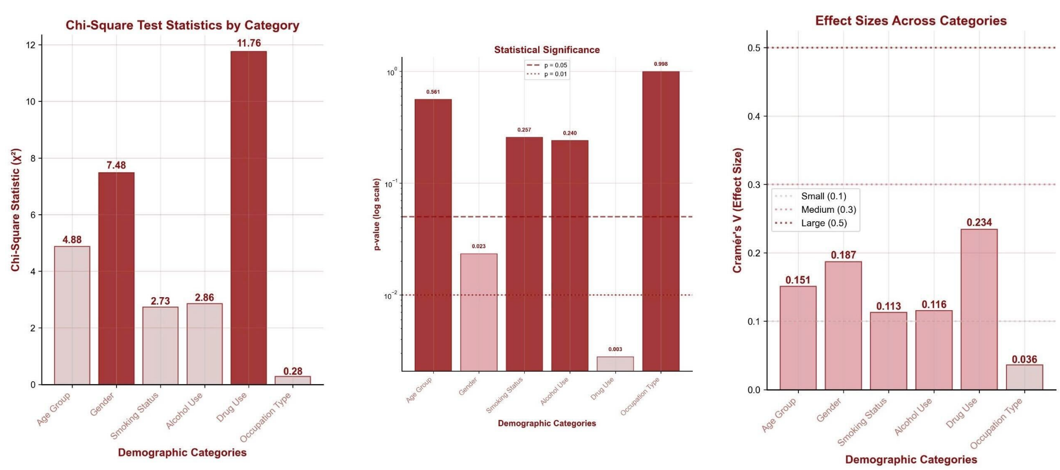

Figure 7: Statistical test results for 6 categories: Chi-square (left), P-values (middle), and Effect sizes (right).

Chi-square tests identified two statistically significant categories (p < 0.05): Gender ($\chi^2 = 7.483$, df = 2, p = 0.0233, V = 0.187) and Drug Use ($\chi^2 = 11.763$, df = 2, p = 0.0028, V = 0.234).

Six categories showed non-significant results (p > 0.05).

All effect sizes were classified as small $(0.1 \leq V < 0.3)$. Total significant categories: 25% (2/8).

B.8

We identify significant but non-critical disparities in model accuracy across age and drug-use subgroups. Accuracy varied systematically by age, with the lowest in the 20–30 cohort (0.500) and the highest in the 50–60 cohort (0.727), yielding a bias magnitude of 0.227 and suggesting that midlife patients align more closely with standard disease patterns and presentation. For drug use, non-users (0.368) performed worse than users (0.714) or unknowns (0.645), producing a disparity of 0.277. Although all disparities fall below critical thresholds, age and drug-use status are clinically meaningful differences and may have an impact on the reliability of diagnosis in practice. The magnitude of fairness in this systematic bias is of significance in obtaining fair model performance on various patient subgroups.

B.9

We observe subgroup disparities across sex, smoking status, and occupation. A lower accuracy in female demographics reflect training set bias as previously established sex differences in the medical predictive models. Differences in occupational groups also indicate that socioeconomic factors can play a part in model stability. These results suggest that the level of bias even at the moderate levels should not be overlooked, as they played a cumulative role across categories to impact the fairness of the population.

B.10 Emergent biases discussion

Our findings indicate that bias in LLM-MAS is structural, with inequities observed across all demographic groups. The absence of parity gaps below 0.1 shows that no subgroup receives unbiased outputs, and the variation in bias severity suggests that demographic complexity interacts with system-level heuristics to amplify disparities. These findings confirm that bias remains a persistent challenge in LLM-MAS and is directly influencing the consistency of its clinical reasoning across demographic groups, and highlight the need for targeted improvements in system design to address fairness.

C Prompting Strategies

C.1 Prompts used in prompting strategy ablation

Table 35: Chain of Thought Prompt

<table>
<tr><th>Role</th><th>Prompt Text</th></tr>
<tr><td>System Prompt</td><td>You are Dr. Agent. Think step-by-step, then output one line that starts with " "'DIAGNOSIS READY:' containing exactly the requested number of diagnoses separated by '|'.</td></tr>
<tr><td>Example Response</td><td>Differential Diagnosis (based on current information): Hepatic causes: Early liver disease (e.g., metabolic storage disorders, early cirrhosis, congenital hepatic fibrosis) even in the absence of jaundice. Protein-losing enteropathy or nephrotic syndrome: Less likely without edema or changes in urine, but still possible. Malabsorption syndromes: Cystic fibrosis, celiac disease (rare in infants), or other congenital GI disorders. - Ascites: Due to hepatic, cardiac, or less commonly, infectious or oncologic causes. Other: Congenital GI anomalies (e.g., Hirschsprung's disease, but usually presents with constipation), or rare metabolic disorders.</td></tr>
</table>

Table 36: Uncertainty Prompt

Role	**Prompt Text**
System Prompt	You are Dr. Agent. Consider diagnostic uncertainty; then output one line that starts with " "'DIAGNOSIS READY:' containing exactly the requested number of diagnoses separated by '|'.
Example Response	Differential Diagnosis at this Stage: Phyllodes tumor: Rapidly growing, firm, mobile mass; can be benign, borderline, or malignant. × Fibroadenoma: Common in younger women, but typically slower growing; can occasionally grow rapidly ("giant fibroadenoma"). Breast carcinoma: Less likely without skin/nipple changes or lymphadenopathy, but cannot be excluded, especially with rapid growth. Other rare tumors: (e.g., sarcoma, lymphoma) less likely but considered if imaging/biopsy is atypical.

Table 37: Differential-First Prompt

Role	**Prompt Text**
System Prompt	"You are Dr. Agent. List the differential internally, then output one line that starts with " "'DIAGNOSIS READY:' containing exactly the requested number of diagnoses separated by '|'."
Example Response	Differential Diagnosis to Consider: - PML (most likely, given risk factors and presentation) - Multiple Sclerosis or other demyelinating disease × (especially if there is a history or MRI features suggestive) - CNS lymphoma** (can mimic PML, especially in immunosuppressed) - CNS vasculitis - Other opportunistic CNS infections (e.g., toxoplasmosis, fungal, viral encephalitis) - Metabolic or toxic leukoencephalopathy

C.2 Results

Table 38: Various prompting strategy effects on metrics. All scenarios were run with each prompting strategy. Default achieved the highest diagnostic accuracy, while other strategies showed marginal differences; process and efficiency metrics were comparable across strategies.

	Diagnostic Accuracy				Diagnostic Process Metrics		Efficiency / Information Quality		
Name	Top-1(%)	Top-3(%)	Top-5(%)	Top-7(%)	Avg dx considered	Avg Tests	Avg Emb	Avg Best Emb	Avg Info Den
Default	56.00	70.67	75.33	77.33	11.30	0.35	0.46	0.87	0.50
Chain-of-Thought	54.67	64.00	68.67	72.00	10.77	0.39	0.47	0.85	0.51
Uncertainty	50.00	64.67	66.67	70.00	10.89	0.35	0.45	0.82	0.51
Differential-first	55.33	64.00	70.67	71.33	11.02	0.37	0.47	0.85	0.50

The default prompting strategy had the highest Top-K accuracy out of all prompting strategies. It also considered the most diagnoses (11.3 average diagnoses considered), and had the highest average embedding score for correct diagnoses (0.87). The differences in between strategy Top-K accuracy is small, with all four strategies having at a Top-1 accuracy of at least 50%.

Average information density scores were very similar across the four different strategies, ranging from 0.50 to 0.51. Both types of embedding similarities are also roughly stable across strategies spanning 0.45-0.47 and 0.82-0.87.

C.2.1 Prompting strategy ablation discussion

Despite being designed to improve reasoning, none of the prompting strategies outperformed the default strategy in Top-K accuracy. However, it is important to consider that default did not outperform the other strategies by large margins; the largest Top-1 accuracy discrepancy between two strategies was 6% (Table 38). Although chain-of-thought reasoning is the second lowest performing in terms of Top-1 accuracy, it has the second-best Top-7 accuracy. This increase could be attributed to the strategy's tendency to generate and follow a larger trail of reasoning, leading the model to produce more plausible diagnoses than other strategies like uncertainty.

The average number of diagnoses considered is the highest for the default strategy at 11.30, and lowest for chain-of-thought and uncertainty at 10.77 and 10.89 respectively (Table 38). With the default strategy considering the most diagnoses and performing the best for Top-1 accuracy, and chain-of-thought and uncertainty considering the least diagnoses but performing the worst for Top-1 accuracy, the question of inefficiency comes to the forefront. While more diagnoses considered means the model is being thorough, it could also lead to less efficient reasoning. This is reflected by the information density score being 0.01 lower for the default strategy when compared to both chain-of-thought and uncertainty (Table 38). While small, this difference translates to real-world clinical scenarios where a doctor would spend more time and resources finding potential diagnoses and ruling out less likely ones, which decreases the efficiency while increasing the overall burden of the diagnostic process on resources.

D Agent structure ablation diagrams

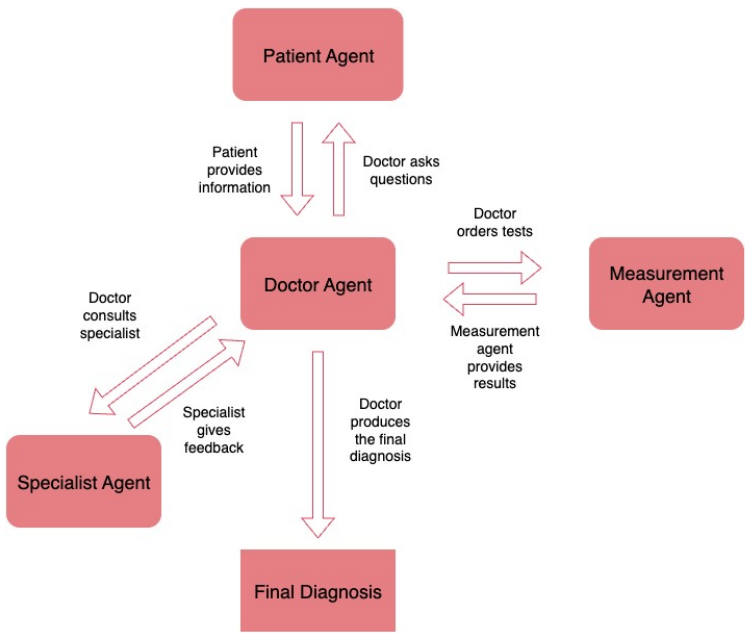

Figure 8: Flow chart of base case agent architecture interactions

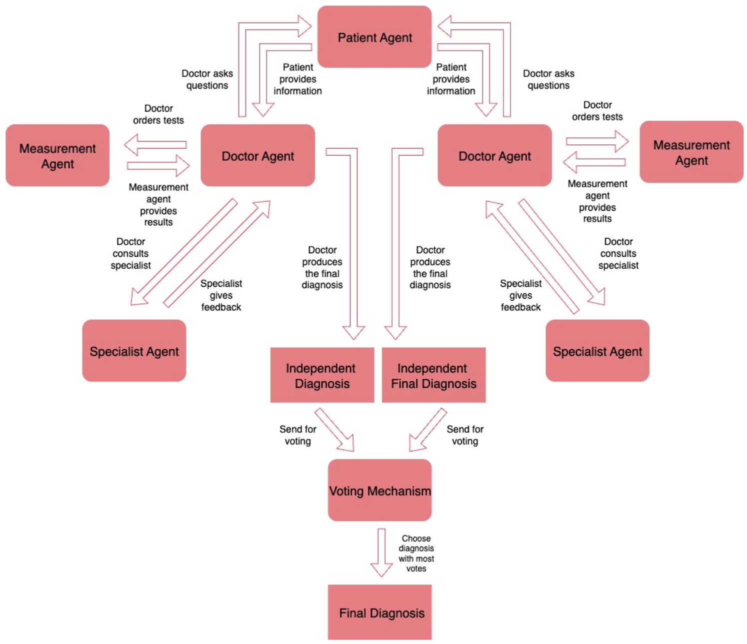

Figure 9: Flow chart of redundant agent architecture interactions

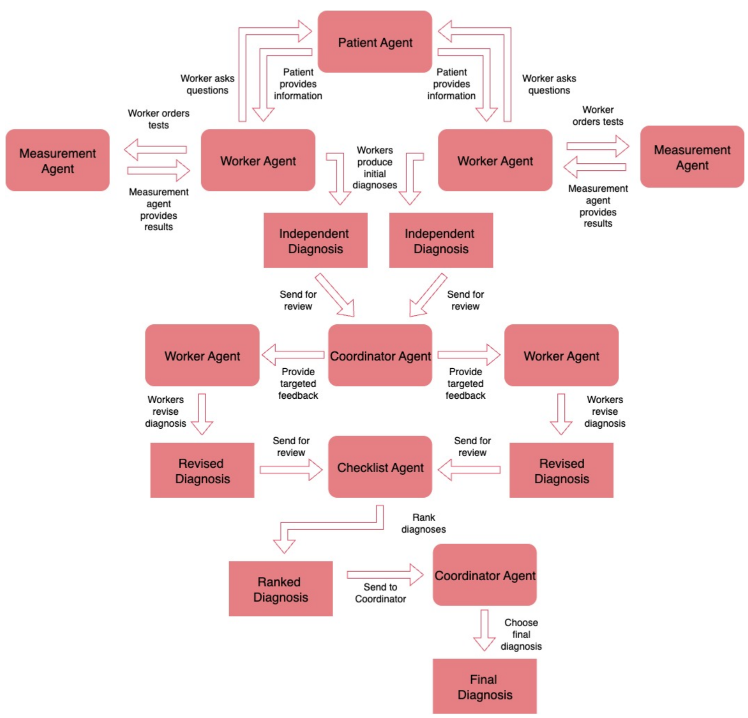

Figure 10: Flow chart of hierarchical agent architecture interactions

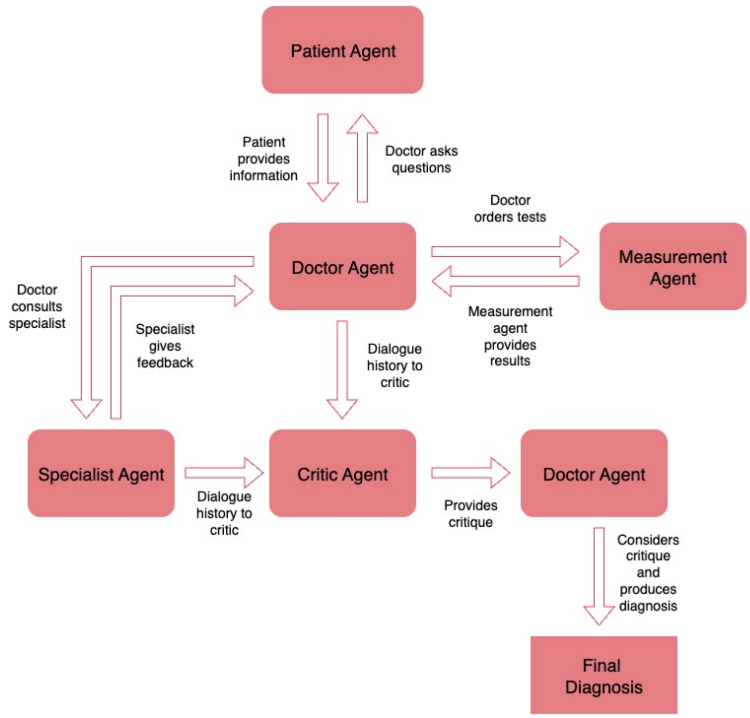

Figure 11: Flow chart of adversarial agent architecture interactions

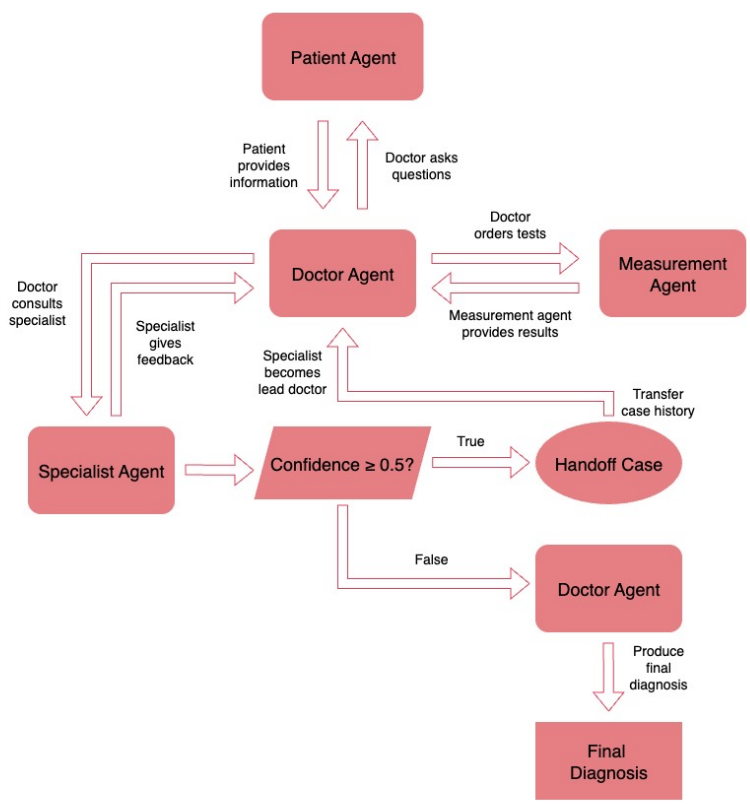

Figure 12: Flow chart of dynamic agent architecture interactions

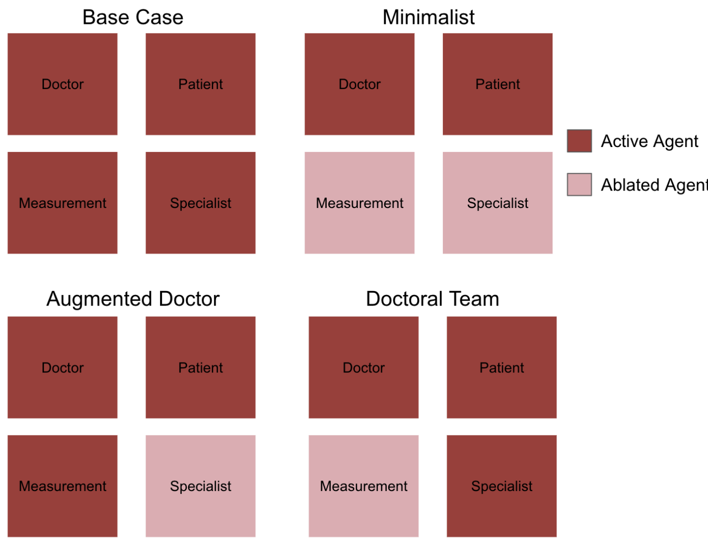

Figure 13: Chart of agents which are ablated from each configuration

E Information quality and efficiency metrics

The average embedding similarity was calculated by taking the cosine similarity between each predicted diagnosis and the ground-truth diagnosis, and then averaging these values across all predictions made by the Doctor Agent. This captures how close the model's predictions are in meaning to the ground truth, helping in cases where the agent outputs a synonymous diagnosis. Likewise, the average best embedding similarity was calculated by averaging the cosine similarity only for predictions judged semantically equivalent to the ground truth. For both of these a score of 1.0 indicates a perfect semantic match.

The information density score was calculated by embedding each Doctor Agent's utterance, computing pairwise cosine similarities, and averaging them to obtain a similarity score s. Information density was defined as 1 – s, where higher values mean that the Doctor Agent's utterances are less redundant and more efficient. This was used to measure how efficient the Doctor Agent is communicating and whether certain systemic changes lead to redundancy in dialogue.