DynamiCare: A Dynamic Multi-Agent Framework for Interactive and Open-Ended Medical Decision-Making

Anonymous ACL submission

Abstract

The rise of Large Language Models (LLMs) has enabled the development of specialized AI agents with domain-specific reasoning and interaction capabilities, particularly in healthcare. While recent frameworks simulate medical decision-making, they largely focus on single-turn tasks where a doctor agent receives full case information upfront—diverging from the real-world diagnostic process, which is inherently uncertain, interactive, and iterative. In this paper, we introduce **MIMIC-Patient**, a structured dataset built from the MIMIC-III electronic health records (EHRs), designed to support dynamic, patient-level simulations. Building on this, we propose **DynamiCare**, a novel **dynamic multi-agent framework** that models clinical diagnosis as a multi-round, interactive loop, where a team of specialist agents iteratively queries the patient system, integrates new information, and dynamically adapts its composition and strategy. We demonstrate the feasibility and effectiveness of DynamiCare through extensive experiments, establishing the **first benchmark** for dynamic clinical decision-making with LLM-powered agents.

1 Introduction

The advent of Large Language Models (LLMs) has laid the foundation for developing specialized AI agents capable of reasoning and interaction tailored to applications in the healthcare domain (Clusmann et al., 2023; Kim et al., 2024b; Saab et al., 2024; Truhn et al., 2024; Zhou et al., 2023). Recent works have leveraged LLMs to simulate medical decision-making (Kim et al., 2024a; Li et al., 2024b; Fan et al., 2024; Jin et al., 2024; Li et al., 2024a; Tang et al., 2023), complemented by the creation of diverse datasets (Jin et al., 2019, 2021; Pal et al., 2022; Chen et al., 2025) designed to mimic real-world medical scenarios and facilitate systematic evaluations.

However, most current AI agents focus on single-turn question-answering tasks (Kim et al., 2024a; Li et al., 2023; Wu et al., 2023; Du et al., 2023; Wang et al., 2022; Tang et al., 2023; Chen et al., 2023). In these scenarios, an LLM-based agent often receives a complete description of an illness at the beginning of the study and is expected to immediately provide a diagnosis. This approach deviates from real clinical practice in two significant ways. First, *the "question" data does not reflect the real diagnostic process*, where patients rarely present complete conditions at the start. Second,*medical diagnosis involves interactive and iterative exchanges*, where healthcare providers progressively elicit relevant details (e.g., symptoms, history, and lab test results) through multiple rounds of interactions.

While some recent studies (Li et al., 2024b; Hu et al., 2024; Schmidgall et al., 2024) have explored interactive diagnostic frameworks with incomplete initial information, they often lack true dynamism—*the ability to adapt the composition and behavior of the agent team based on newly acquired information*. In contrast, real clinical environments necessitate continuous, context-aware adjustments, including modifications to the healthcare team structure in response to evolving patient needs and clinical complexity.

To address these limitations, we propose **MIMIC-Patient** (the left part of Figure 1), a patient-level dataset that compiles diverse medical information for each patient based on the MIMIC-III Clinical database (Johnson et al., 2016), and **DynamiCare** (the right part of Figure 1), a dynamic, interactive framework for simulating clinical decision-making. DynamiCare consists of a Patient System and a Doctor System, operating in a six-step loop: 1) Initialization: create a visit log using basic patient information from MIMIC-Patient; 2) Team Formation: a central agent recruits specialist agents based on the visit log; 3) Special-

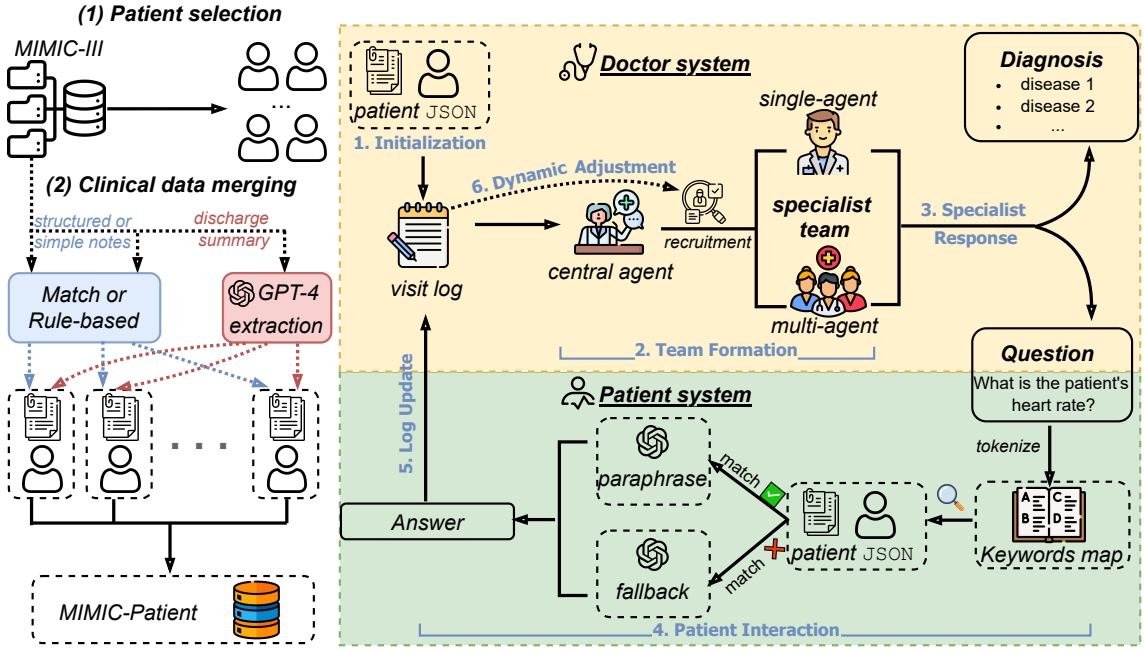

Figure 1: Illustration of the MIMIC-Patient dataset and DynamiCare framework. **Left**: the construction process of MIMIC-Patient. **Right**: The DynamiCare framework, which consists of a Patient System and a Doctor System, operating in a six-step loop: 1) initialization; 2) team formation; 3) specialist response; 4) patient interaction; 5) log update; 6) dynamic adjustment.

ist Response: the team generates a diagnosis or follow-up question; 4) Patient Interaction: the Patient System answers the question using the patient record; 5) Log Update: the Q&A pair is added to the visit log; 6) Dynamic Adjustment: the central agent updates the team based on the new information, and the loop continues until a diagnosis is made or reaches the interaction round limit.

Our contributions are summarized as follows:

- We establish the **MIMIC-Patient** benchmark, a patient-centric benchmark dataset derived from MIMIC-III, which structures diverse medical information to support interactive and open-ended decision-making tasks.
- We introduce **DynamiCare**, a **novel multi-agent framework** to model clinical reasoning as a dynamic, interactive process that can adapt its structure and strategy based on newly acquired patient information.
- We conduct extensive experiments using DynamiCare on MIMIC-Patient, establishing a **benchmark** that set a foundation for future research on dynamic medical agents.

2 Related Works

Clinical Dataset Recent works (Singhal et al., 2023, 2025; Nori et al., 2023; Liévin et al., 2024) have produced a wide range of benchmark datasets to evaluate the performance of LLMs in the clinical domain. Notable examples include MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022), which feature question-answer pairs derived from medical licensing exams or biomedical literature. Additionally, several works (Pampari et al., 2018; Fan, 2019; Kweon et al., 2024) have explored electronic health record (EHR) datasets. For example, Kweon et al. (2024) introduced EHRnoteQA, a QA dataset derived from discharge summaries in MIMIC-IV (Johnson et al., 2023). Although valuable for assessing the factual medical knowledge of LLMs, these datasets only include predefined questions and answers without the ability to support the dynamic, interactive modeling of real clinical encounters. To address this, Li et al. (2024b) proposed MedIQ, an interactive benchmark for medical evaluation. However, the construction is based on MedQA and Craft-MD (Johri et al., 2024), which lacks the complexity of real-world clinical scenarios. In practice, clinical tasks

often involve extracting, synthesizing, and reasoning over heterogeneous patient information—such as clinical notes, laboratory results, and timelines of care—which these benchmarks do not often capture.

LLM Agents in Medical Decision Making A growing body of research (Zhou et al., 2023; Wang et al., 2024) has proposed frameworks that leverage single or multiple LLM agents to support medical decision-making. These approaches typically rely on prompt engineering to guide LLMs in completing clinical tasks. Notable paradigms include role-playing (Li et al., 2023; Wu et al., 2023), debate (Du et al., 2023; Liang et al., 2023), voting (Wang et al., 2022), multi-disciplinary collaboration (Tang et al., 2023), and group discussion (Chen et al., 2023). While these agentic frameworks have demonstrated performance improvements in specific settings, they are often built around a fixed and pre-defined set of agents, making them inherently static. To address this rigidity, Kim et al. (2024a) introduced MDAgents, a more adaptive framework that employs a complexity assessment to determine the appropriate number of agents for a given task. However, MDAgents can fall short of capturing the dynamic nature of real-world clinical workflows, where the composition and involvement of specialist agents should evolve over the course of multi-turn interactions between clinicians and patients.

3 MIMIC-Patient

We build MIMIC-Patient, a dataset derived from the MIMIC-III database (Johnson et al., 2016). MIMIC-III is a large, publicly available dataset containing de-identified clinical data from over 40,000 patients admitted to critical care units at the Beth Israel Deaconess Medical Center between 2001 and 2012. However, it can be extremely challenging for our dynamic scenarios given its inherent complexity of clinical records. For instance, individual patient admission may include hundreds of diagnoses—far exceeding the input capacity of current LLMs. Moreover, the clinical data are distributed across multiple relational tables, complicating effective integration and interpretation. To address these challenges, we adopt a two-stage data processing approach (the left part of Figure 1) to construct a more structured and analysis-friendly dataset called MIMIC-Patient.

Data Type	Clinical Info
Structured Data	Admission Info Demographics Diagnoses Prescription Procedure [1] Chart Data Lab Data
Semi-Structured Text	ECG reports [2] Echo reports [3] Radiology reports
Unstructured Text	Discharge Summary

Table 1: Clinical information included in the patient JSON.[4]

Data selection MIMIC-III contains 58,976 hospital admissions, each representing a unique hospital stay with associated clinical data. We begin by selecting admissions that met all three criteria:

- The admission has fewer than five diagnosed diseases. In the MIMIC-III dataset, some admissions have a large number of multimorbidities, which can introduce excessive complexity and noise.
- The admission is neither for a newborn nor deceased. Newborns often follow different clinical pathways and have distinct data structures, while deceased patients may have incomplete or atypical medical records that could bias the results.
- The admission has sufficient clinical data available, i.e., at least containing all structured data and a discharge summary, to ensure that both the doctor and patient agents have access to meaningful and informative input for decision-making.

After filtering for valid admissions, 2,597 remained. Note that a single patient may have multiple admissions. To ensure a one-to-one mapping between patients and admissions, we retain only one admis-

[1] Clinical procedures performed, with codes and timestamps.

[2] Electrocardiogram reports and interpretations.

[3] Echocardiogram results, including measurements like ejection fraction.

[4] More details of the database can be found at: https://mimic.mit.edu/docs/iii/tables/

sion per patient, resulting in 2,452 unique admissions. From this subset, we then randomly select 500 admissions. Since each admission corresponds to a unique patient in a one-to-one manner, we refer to our dataset as patient-level and use the term "patient" interchangeably with the corresponding admission in the following descriptions.

Clinical data merging In the MIMIC-III database, clinical data are distributed across multiple tables. To centralize and structure this information, we extract and merge the relevant data for each patient into a single `JSON` file (see examples in Appendix A). For structured data (e.g., charted observations), we directly match and integrate the entries based on patient IDs. For unstructured data (e.g., free-text clinical notes), we employ two extraction methods. First, we develop a *rule-based* approach to extract information from semi-structured text—such as the `Impression` and `Findings` sections in radiology reports—by identifying and normalizing known section headers. Second, for more complex unstructured texts like discharge summaries, we leverage GPT-4 (Achiam et al., 2023) to parse the content and generate a structured `JSON` representation.

As a result, the final dataset comprises 500 `JSON` files—one for each patient—containing a variety of clinical records, from demographic details to lab results. A complete list of the included clinical information is provided in Table 1.

4 DynamiCare

To enable interactive medical decision-making, DynamiCare (the right part of Figure 1) is designed to comprise two components: the **Patient System** and the **Doctor System**. At a high level, the Patient System responds to queries from the Doctor System, which in turn updates its strategy based on the received information. In this section, we first introduce the Patient System, followed by the Doctor System. We then describe the overall workflow and explain how DynamiCare facilitates dynamic interactions.

4.1 Patient System

To enable interactive querying of patient information by a doctor agent (e.g., an LLM), we build a Patient System capable of responding to natural language questions using structured patient records. Importantly, doctor queries vary in complexity—some map directly to specific data fields, while others require interpretation and integration across multiple record types. In order to ensure the reliability of its responses and minimize LLM hallucinations, we propose a two-stage answering process (the green part of Figure 1) that combines rule-based keyword matching with fallback LLM inference.

For each query, the Patient System first extracts relevant keywords using regular expression-based tokenization. These keywords are then mapped to specific sections of the patient's structured `JSON` file—such as demographics, prescriptions, or lab results—based on a predefined keyword-to-section mapping dictionary (example in Appendix A). Once the relevant section is identified, the Patient System retrieves the corresponding information and uses GPT-4.1 (OpenAI, 2025) to formulate a response in the patient's voice.

To handle cases where no section is matched or the retrieved information is ambiguous or missing, the system falls back to direct LLM inference. In such scenarios, the patient `JSON` (de-identified, removing admission and demographics) is provided as context to GPT-4.1, which then generates an answer based on the broader clinical context.

The generated answer is then passed to the Doctor System to support real-time adaptation of its composition and strategy.

4.2 Doctor System

Previous studies (Kim et al., 2024a; Gilboy et al., 2012; Christ et al., 2010; Wuerz et al., 2000) have shown that flexible adaptation of specialist teams to varying patient conditions is both effective and promising for multi-agent system design. Moreover, as the doctor agent actively guides the interaction to gather additional information, the patient's condition representation may evolve. This highlights the need for a dynamic Doctor System capable of adjusting its composition in real time.

Dynamic Specialist Team In DynamiCare, the Doctor System (the yellow section of Figure 1) incorporates a **Central Agent** that acts as a senior medical coordinator, dynamically managing a **Specialist Team** based on the evolving context of each patient case. The Central Agent continuously reviews the visit log—which captures previous rounds of question-answer interactions—and decides whether to update the composition of the Specialist Team. This dynamic process ensures that the team remains aligned with the most cur-

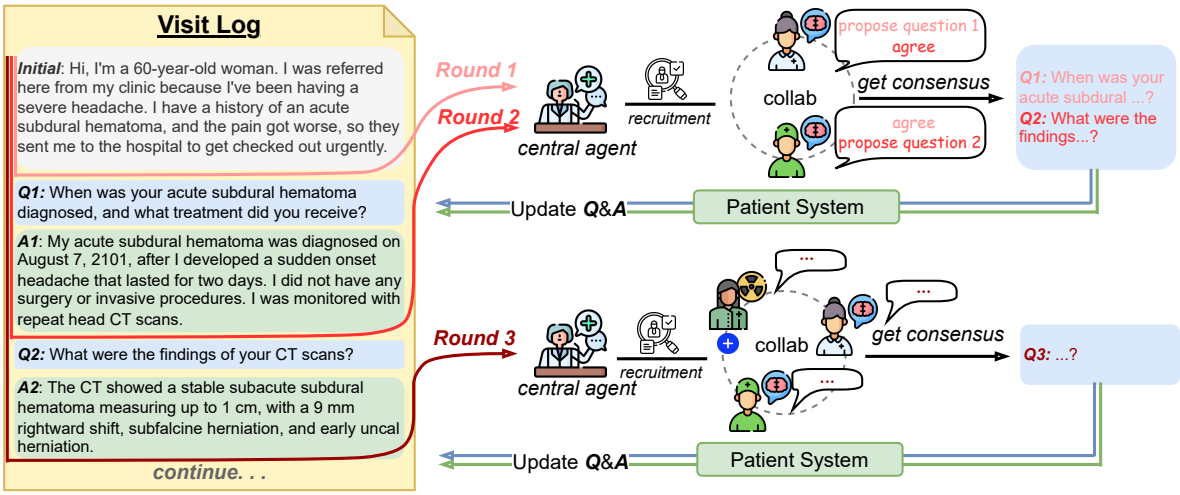

Figure 2: Illustrative example of DynamiCare with a multi-specialist team. The figure presents three rounds of interaction, where in each round the specialist team poses a question to the patient. In Rounds 1 and 2, the team consists of a neurologist and a neurosurgeon. Notably, in Round 3, the Central Agent recruits an additional specialist—a radiologist—to assist with analyzing the patient's CT scans.

rent information, adding or removing specialists as needed.

Collaborative Decision Protocols For simpler cases, the Central Agent may assign a single specialist, while more complex scenarios may involve multiple specialists with complementary expertise. The specialist team then decides whether to issue a diagnosis or request additional information. *In the single-specialist setting*, the agent analyzes the visit log, evaluates diagnostic confidence, and either provides a diagnosis (if confidence is high) or asks a follow-up question (if confidence is low). *In the multi-specialist setting*, each agent independently proposes a response (diagnosis or question) along with a confidence score. Specialists then vote on each other's proposals; the first response to meet a predefined agreement threshold is accepted. If no consensus is reached, the response with the highest confidence is selected. This adaptable framework ensures that diagnostic reasoning is both context-aware and continuously updated in response to new patient information.

4.3 Dynamic Interaction Workflow

The interaction between the Patient and Doctor Systems follows a dynamic, iterative workflow that enables real-time information gathering and diagnosis. Figure 2 gives a detailed illustration with a concrete example for this workflow. The process proceeds in six steps:

1. **Initialize Visit Log**: A *visit log* is created using the patient's basic information, which contains demographics, basic relevant symptoms, and the reason for the visit.

2. **Specialist Team Formation**: The *Central Agent* receives the initial visit log. Based on this input, the Central Agent initializes an appropriate *Specialist Team* tailored to the current case complexity.

3. **Specialist Response Generation**: The Specialist Team analyzes the current visit log and, following the protocols described in Section 4.2, generates a response. This response can be either a follow-up question or a diagnostic conclusion.

4. **Patient Interaction**: If the response is a diagnosis, it is treated as the final decision, and the interaction ends. If the response is a question, the *Patient System* is queried to retrieve an appropriate answer.

5. **Update Visit Log**: The current question-answer pair is appended to the visit log to maintain a comprehensive record of the evolving case.

6. **Dynamic Adjustment**: The Central Agent re-evaluates the current visit log and the performance of the existing specialist team. Based on new information or shifting diagnostic needs, it may revise the team composition

before returning to Step 3 for next-round analysis.

Additionally, to ensure termination, a manual stopping criterion is defined: if the system reaches a predefined maximum number of interaction rounds without a conclusive diagnosis, the current specialist team is prompted to generate the best possible diagnosis with the available information, thereby concluding the session. All prompts are detailed in AppendixB.

5 Experiments and Results

5.1 Experimental Setup

We evaluate both of our proposed Doctor System and its single-agent variant in DynamiCare on the MIMIC-Patient dataset (500 patient records) using GPT-4.1 and GPT-4o-mini. In the single-agent setting, the multi-agent specialist team is disabled and only one agent performs diagnostics per round. GPT-4.1 is also used in the Patient System for answer generation. For comparison with external baselines, we test our model on the MEDIQ benchmark(Li et al., 2024b), using all 140 samples from iCraftMD and 200 randomly selected cases from iMedQA, following MEDIQ's interactive multiple-choice setting and using GPT-4.1 for all runs.

5.2 Experimental Results

Table 2 presents the performance of our proposed Doctor System, including its multi- and single-agent variants (with the specialist team fixed to the single-agent version). The backbone models use GPT-4.1 and GPT-4o-mini. The results demonstrate the effectiveness of the dynamic multi-agent setting across multiple metrics and model variants.

We then evaluate the results using the top-k hit rate (Hit@5, Hit@10) [5] and recall (Rec@5, Rec@10) [6], which measure the system's ability to rank correct diagnoses among its predictions. As the doctor agent is prompted to return a list of up to 10 likely diagnoses, and the ground truth typically includes between 1 to 5 correct labels, Hit@K metrics help assess how well the system captures correct diagnoses within a ranked list, whereas Rec@K offers a perspective on how many ground truth diagnoses are successfully identified.

To standardize diagnosis terminology and enable consistent comparison, we map all predicted diagnoses to ICD-9 codes using the BioPortal API (Whetzel et al., 2011). We consider a prediction correct if the first three digits of the predicted ICD-9 code match any of the ground truth codes. This reflects a clinically meaningful level of accuracy, as the first three digits of an ICD-9 code typically represent the high-level diagnostic category. (Rajkomar et al., 2018; Shi et al., 2017). It is worth noting that, even under this relatively relaxed criterion, the classification task remains extremely challenging, with over 1,000 possible 3-digit ICD-9 code groups in the prediction space.

From Table 2, we observe that the proposed dynamic multi-agent system consistently outperforms the single-agent variant across all metrics and LLM models. These improvements indicate that dynamically adjusting the specialist team based on patient complexity—as implemented in the multi-agent setting—enhances diagnostic performance, especially in our setting involving open-ended medical diagnosis with rich and often ambiguous patient information.

Additionally, model quality plays a significant role. GPT-4.1 consistently outperforms GPT-4o-mini across all settings, emphasizing the importance of a strong language model backbone in achieving accurate medical reasoning. This is further supported by Ave-Q, the average number of questions asked per patient case, where we observe more questions asked with stronger models and more complex settings, reflecting a more interactive diagnostic process.

5.3 Cross Comparison with MEDIQ

To further validate the effectiveness of our dynamic multi-agent setting, we conduct additional evaluations on the MEDIQ interactive benchmark(Li et al., 2024b), which includes two interactive QA tasks: iMedQA and iCraftMD. These tasks are framed as multiple-choice question answering (MCQ), which is relatively simpler than our open-ended diagnosis generation task. For evaluation, we use the full iCraftMD dataset (140 samples) and randomly select 200 cases from iMedQA. Prompt templates for our Doctor System are appropriately adapted to match the multiple-choice format, all experiments use GPT-4.1. As shown in Table 3, our Doctor System significantly outperforms the best MEDIQ baseline in both datasets, further demonstrating the capability of our model.

[5] Hit@K measures whether at least one of the matched (ground truth) items appears in the top K results returned by the model.

[6] Recall@K measures the proportion of all matched items that are found in the top K results.

Agent	GPT Version	Hit@5	Hit@10	Rec@5	Rec@10	Ave-Q
Multi	GPT-4.1	**63.4**	**71.6**	**43.2**	**58.8**	7.55
	GPT-4o-mini	51.4	58.6	30.2	43.2	2.78
Single	GPT-4.1	58.0	63.2	31.0	41.7	3.83
	GPT-4o-mini	47.8	54.8	24.9	31.8	0.74

Table 2: Performance comparison between the proposed multi-agent Doctor System and a single-agent variant across two LLMs (GPT-4.1 and GPT-4o-mini). Metrics include Hit@K, Rec@K, and Ave-Q (average number of questions asked).

Agent	Dataset	Accuracy
MedIQ	iMedQA	67.0
	iCraftMD	72.1
DynamiCare	iMedQA	**92.0**
	iCraftMD	**96.4**

Table 3: Accuracy comparison on multiple-choice question answering tasks from the MEDIQ benchmark, across two datasets (iMedQA and iCraftMD). Our dynamic multi-agent system (DyamiCare) outperforms the MedIQ baseline under the same GPT-4.1 setup.

5.4 Patient System Evaluation

Metric	A	B	C	Average
Truthfulness	1.99	1.95	1.92	1.95
Relevance	1.84	1.75	1.79	1.79

Table 4: Manual evaluation of simulated patient responses across 100 patient sessions. Each response was rated by three annotators on Truthfulness and Relevance using a 3-point scale (0–2). The table reports the average scores per annotator and the overall mean across annotators.

To assess the quality of the responses generated by our Patient System, we conduct a manual evaluation of 100 randomly selected patients' question-answer history (multi-agent setting, GPT-4.1). Each Q&A log is independently annotated by three medical students along two dimensions: Truthfulness and Relevance.

- **Truthfulness** measures whether the patient's response is consistent with his/her JSON record, using a 3-point scale: 0 = incorrect, 1 = partially correct, 2 = fully correct.
- **Relevance** assesses whether the answer directly and adequately responds to the doctor's question, also using a 3-point scale: 0 = irrelevant, 1 = somewhat related, 2 = highly relevant.

The scores from the three annotators were averaged across all questions per patient and reported separately for each annotator (A, B, C), along with the overall mean as shown in Table 4. This evaluation provides a quantitative estimate of the reliability and contextual appropriateness of the patient agent's responses, ensuring that it supports trustworthy and contextually appropriate interactions throughout the diagnostic process.

5.5 Disease Case Study

We observe substantial variation in diagnostic performance across ICD-9 classes, with top-5 accuracy ranging from 32% to 100% (mean: 63.4%), and top-10 accuracy from 40% to 100% (mean: 71.6%). Excluding classes with insufficient sample size, high-performing categories include diagnoses like musculoskeletal diseases and circulatory system disorders, whereas lower performance is seen in neoplasms, external causes of injury, and symptoms or ill-defined conditions.

These differences appear to stem from three main factors. First, **case complexity** plays a key role (Khan and O'Sullivan, 2024; McDuff et al., 2025). Conditions with well-defined, localized symptoms—such as musculoskeletal or circulatory disorders—tend to involve fewer comorbidities and more predictable clinical patterns, enabling the model to identify the correct diagnosis with fewer reasoning steps. In contrast, complex conditions like neoplasms or non-specific symptom clusters often involve overlapping or ambiguous presentations that require integration of contextual or longitudinal information, making accurate diagnosis considerably more challenging within a constrained Q&A framework.

Second, **disease prevalence and representa-**

ICD-9 codes	Definition	Hit@5	Hit@10	Sample Size
630-679	complications of pregnancy, childbirth, and the puerperium	100	100	1
760-779	certain conditions originating in the perinatal period	85.71	85.71	7
710-739	diseases of the musculoskeletal system and connective tissue	76.00	76.00	25
390-459	diseases of the circulatory system	74.09	81.73	301
740-759	congenital anomalies	69.23	76.92	26
240-279	endocrine, nutritional and metabolic diseases, and immunity disorders	67.27	77.58	165
520-579	diseases of the digestive system	63.77	72.46	69
460-519	diseases of the respiratory system	62.90	72.58	62
320-389	diseases of the nervous system and sense organs	61.22	71.43	49
290-319	mental disorders	57.97	69.57	69
280-289	diseases of the blood and blood-forming organs	55.56	75.00	36
800-999	injury and poisoning	50.56	59.55	89
580-629	diseases of the genitourinary system	50.00	66.67	12
680-709	diseases of the skin and subcutaneous tissue	50.00	50.00	6
780-799	symptoms, signs, and ill-defined conditions	48.78	68.29	41
E and V codes	external causes of injury and supplemental classification	48.28	57.93	145
140-239	neoplasms	32.00	40.00	50
001–139	infectious and parasitic diseases	-	-	0

Table 5: Diagnostic accuracy across ICD-9 code categories. Each row represents a high-level disease group, along with its corresponding accuracy and sample size. Categories with accuracy above the overall mean (Hit@5: 63.4%, Hit@10: 71.6%) are highlighted in green. Sample size indicates the number of patient–diagnosis instances assigned to each ICD-9 category; since each patient may have multiple codes, the total exceeds the number of unique patients.

tion may influence GPT's performance (Sandmann et al., 2024). Although the system is not fine-tuned, its base knowledge reflects conditions that are more frequently discussed in medical literature and clinical practice. As a result, the model tends to perform better on prevalent diseases with well-defined patterns (e.g., cardiovascular and circulatory diseases).

Finally, the **disease-specific nature of Q&A interactions** influences diagnostic success. Diseases with clear symptomatology allow for efficient question targeting, enabling quick narrowing the differential diagnosis. Conversely, vague or multi-system conditions—such as those classified under neoplasms or ill-defined symptoms—require broader questioning and yield more diffuse information, often resulting in longer differential lists where the correct diagnosis may be deprioritized. Additionally, categories requiring contextual or social information (e.g., external causes of injury) are more likely to be missed when such details are not explicitly elicited.

6 Discussion

This study introduces a dynamic multi-agent framework for clinical decision-making, grounded in a realistic simulation of the diagnostic process. Unlike prior works that rely on single-turn or static multi-agent setups, our system dynamically adapts its specialist composition in response to evolving patient information. Experiments on the MIMIC-derived open-ended diagnosis benchmark demonstrate that our model consistently outperforms a single-agent variant, particularly when reasoning over complex and ambiguous cases. Moreover, our patient system was shown to produce responses that are both truthful and relevant, supporting a coherent and trustworthy dialogue process.

7 Limitations

While our approach shows promising results, several limitations remain, opening opportunities for future research and improvement. First, our system currently operates on textual and tabular data only. Incorporating other clinically significant modalities such as medical imaging, genomics, or sensor data could enable richer and more accurate diagnostic reasoning. Second, real patients may volunteer important information even if not directly asked. Future versions of the patient system could be designed to simulate such proactive behavior, improving the realism of interactions. Third, while our dynamic framework already improves reasoning, further enhancement may come from integrating retrieval-augmented generation (RAG), external medical knowledge bases, or modular expert components fine-tuned for specific specialties.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. 2025. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3563–3599.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Michael Christ, Florian Grossmann, Daniela Winter, Roland Bingisser, and Elke Platz. 2010. Modern triage in the emergency department. *Deutsches Ärzteblatt International*, 107(50):892.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Jungwei Fan. 2019. Annotating and characterizing clinical sentences with explicit why-qa cues. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 101–106.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv e-prints*, pages arXiv–2402.

Nicki Gilboy, Paula Tanabe, Debbie Travers, Alexander M Rosenau, and 1 others. 2012. Emergency severity index (esi): a triage tool for emergency department care, version 4. *Implementation handbook*, 2012:12–0014.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. *arXiv preprint arXiv:2402.03271*.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, W John Wilbur, Zhe He, Andrew Taylor, Qingyu Chen, and 1 others. 2024. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *arXiv preprint arXiv:2402.13225*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Maria Palwasha Khan and Eoin Daniel O'Sullivan. 2024. A comparison of the diagnostic ability of large language models in challenging clinical cases. *Frontiers in Artificial Intelligence*, 7:1379297.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024a. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024b. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwang Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of

large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and 1 others. 2024a. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. 2024b. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888.

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

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed May 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, and 1 others. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Sarah Sandmann, Sarah Riepenhausen, Lucas Plagwitz, and Julian Varghese. 2024. Systematic analysis of chatgpt, google search and llama 2 for clinical decision support tasks. *Nature Communications*, 15(1):2050.

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

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. *arXiv preprint arXiv:1711.04075*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

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

Daniel Truhn, Jan-Niklas Eckardt, Dyke Ferber, and Jakob Nikolas Kather. 2024. Large language models and multimodal foundation models for precision oncology. *NPJ Precision Oncology*, 8(1):72.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark A Musen. 2011. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(suppl_2):W541–W545.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Richard C Wuerz, Leslie W Milne, David R Eitel, Debbie Travers, and Nicki Gilboy. 2000. Reliability and validity of a new five-level triage instrument. *Academic emergency medicine*, 7(3):236–242.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, and 1 others. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Patient System

Patient JSON (example)


```
"Patient": {
    "Admission_info": {
      "patient_id": ****,
      "admission_id": ****,
      "admission_diagnosis": "arrhythmia"
    },
    "Demographics": {
      "insurance": "private",
      "language": "engl",
      "marital_status": "married",
      "ethnicity": "white",
      "gender": "M",
      "age": 60
    },
    "Diagnoses": [[
        "4019",
        "Hypertension NOS",
        "Unspecified essential hypertension"
      ],...],
    "Prescription": [
      "Sodium Chloride 0.9% Flush",
      "Lisinopril",
      "Heparin",...],
    "Introduction": "Hi, I'm a 60-year-old male. I was referred here by my clinic ...my doctor
         was concerned about a possible arrhythmia.",
    "ECG": [[
        "2105-03-03",
        "Atrial pacing and A-V conduction which is new compared to previous tracings."
      ],...],
    "Radiology": [{
        "time": "2105-03-03",
        "part": "CHEST (PA & LAT)",
        "medical condition": "60 year old man with new dual chamber",
        "ppm reason for this examination": "Evaluate lead position",
        "final report history": "Pacemaker placement.",
        "findings": "In comparison with the study of , there has been placement of ... other
            acute cardiopulmonary disease."
      }...],
    "Allergies": "No Known Allergies / Adverse Drug Reactions",
    "Chief Complaint": "Fatigue, lightheadedness, bradycardia, sinus pauses",
    "Major Surgical or Invasive Procedure": "Pacemaker placement (St. Medical Accent PM2210
        dual chamber pacemaker)",
    "Physical Exam": { "Admission": {
        "VS": "T=98.0 BP=158/91 HR=61 RR=18 O2 sat=95",
        "General": "WDWN M in NAD. Oriented x3. Mood, affect appropriate. Fit",
        "HEENT": "NCAT. Sclera anicteric. PERRL, EOMI. Conjunctiva were pink, no pallor or
            cyanosis of the oral mucosa. No xanthalesma.",
        "Neck": "Supple with JVP below clavicle at 90 degrees.",
        "Cardiac": "PMI located in 5th intercostal space, midclavicular line. RR, normal S1, S2
            . No m/r/g. No thrills, lifts. No S3 or S4.",
        "Lungs": "No chest wall deformities, scoliosis or kyphosis. Resp were unlabored, no
            accessory muscle use. CTAB, no crackles, wheezes or rhonchi.",
        "Abdomen": "Soft, NTND. No HSM or tenderness. Abd aorta not enlarged by palpation. No
            abdominial bruits.",
        "Extremities": "No c/c/e. No femoral bruits.",
        "Skin": "No stasis dermatitis, ulcers, scars, or xanthomas.",
        "Pulses": {
          "Right": "Carotid 2+ Femoral 2+ Popliteal 2+ DP 2+ PT 2+",
          "Left": "Carotid 2+ Femoral 2+ Popliteal 2+ DP 2+ PT 2+"}
      },...},
    "Respiratory": { "O2 saturation pulseoxymetry": [ ["2105-02-28 03:15:00", "97.0 %"
        ],...],...}
```

Keyword Mapping Dictionary (example)

```
keyword_mapping = {
    # Demographics
    ('age',): ('Demographics', 'age'),
    ('language', 'english', 'spanish'): ('Demographics', 'language'),
    ('religion', 'religious'): ('Demographics', 'religion'),
    ('marital', 'married', 'single', 'divorced', 'widowed'): ('Demographics', 'marital_status'
        ),
    ('gender', 'sex'): ('Demographics', 'gender'),
    ('insurance',): ('Demographics', 'insurance'),
    ('ethnicity',): ('Demographics', 'ethnicity'),
    # Medications
    ('prescription', 'medications', 'medication', 'drugs'): ('Prescription', None),
    ('admission medications', 'initial meds'): ('Medications on Admission', None),
    # Procedures
    ('procedure', 'surgery', 'operation'): [('Procedure', None), ('Major Surgical or Invasive
        Procedure', None)],
    # Imaging and reports
    ('electrocardiogram', 'ecg'): ('ECG', None),
    ('echocardiogram', 'echo'): ('Echo', None),
    ('radiology', 'x-ray', 'ct', 'mri', 'imaging'): ('Radiology', None),
    # History
    ('hpi', 'present illness', 'history of present illness'): ('History of Present Illness',
        None),
    ('past medical', 'pmh', 'past medical history'): ('Past Medical History', None),
    ('family history',): ('Family History', None),
    ('social history', 'drinking', 'smoking', 'drug use', 'tobacco', 'alcohol'): ('Social
        History', None),
    # Allergies
    ('allergy', 'allergies', 'allergic'): ('Allergies', None),
    # Physical exam
    ('heent', 'head', 'eyes', 'ears', 'nose', 'throat'): ('Physical Exam.Admission', 'HEENT'),
    ('physical exam',): ('Physical Exam.Admission', None),
    ...}
```

B Doctor Agent on MIMIC Open-Ended Diagnosis

Initial Specialist Triage Prompt

```
You are a general practitioner triaging a new patient. Based on the patient's initial
    admission information, recommend one or more medical specialists to consult.
The maximum number of specialists is 5.
Return your answer in the following JSON format only:
{
"RATIONALE": "<short justification>",
"SUGGEST_SPECIALISTS": [<list of specialists>]
}
```

Central Agent Coordination Prompt

```
You are a central medical coordinator overseeing a diagnostic team.
Based on the current patient case and the specialists already involved, decide if additional
    experts are needed or if any can be removed.
Only suggest adding new specialists if there's missing domain knowledge. Only suggest removing
     specialists if their role has been fully covered or no longer needed.
The maximum number of specialists is 5.
Respond in this JSON format only:
{
"ADD": [<specialists to add>],
"REMOVE": [<specialists to remove>],
"UPDATED_LIST": [<updated specialists>],
"RATIONALE": "<short justification>"
}
```

Specialist Confidence Rating Prompt

```
You are a {spec}. Based on the current patient case, decide whether you are confident enough
    to make a diagnosis or whether more information is needed.

Choose between the following ratings:
"Very Confident"- The diagnosis is strongly supported by current information, and no major
    uncertainties remain.
"Somewhat Confident"- The diagnosis is likely given the evidence, but a bit more information
    would increase certainty.
"Neither Confident or Unconfident"- Some clues suggest a possible diagnosis, but key details
    are still missing.
"Somewhat Unconfident"- Several diagnoses remain plausible; more data is needed to narrow them
     down.
"Very Unconfident"- There is too little evidence to form a reasonable diagnostic opinion.

Respond in the following format only:
DECISION: chosen rating from the above list.
```

Solo Specialist Diagnosis or Question Prompt

```
You are a {spec}. Based on the current patient case, list the top 10 most likely diagnoses for
     this patient.
Only include the **diagnosis name** in the list.
Respond in this JSON format only:
{
"RESPONSE_TYPE": "diagnosis",
"RESPONSE_CONTENT": "[<your diagnosis list>]",
"RATIONALE": "<brief justification>"
}
```

Solo Specialist Follow-Up Question Prompt

```
You are a {spec}. Based on your medical expertise and the current information, propose the
    most important next question that would help you narrow down or confirm a diagnosis.
The question should be specific and relevant to the case.
Do not repeat any questions from the previous conversation log or ask about information
    already provided.
Avoid asking about topics already answered with "I don't know" or "not in chart"
If referencing labs, vitals, ECG, radiology, etc. - which may have multiple time points - be
    clear about the desired time window.
Respond in this JSON format only:
{
"RESPONSE_TYPE": "question",
"RESPONSE_CONTENT": "<your follow-up question>",
"RATIONALE": "<brief justification>"
}
```

Collaborative Decision Proposal Prompt

```
You are a {spec}. Based on your medical expertise and the current information, respond with
    either:
- A list of the top 10 most likely diagnoses (if you believe you can make a clinical judgment
    now), or
- A follow-up question (if you believe more information is needed to proceed)

Only include the **diagnosis name** in the list.
The question should be specific and relevant to the case.
Do not repeat any questions from the previous conversation log or ask about information
    already provided.
Avoid asking about topics already answered with "I don't know" or "not in chart"
If referencing labs, vitals, ECG, radiology, etc. - which may have **multiple time points** -
    be clear about the **desired time window**.

Respond in this JSON format only:
{
"RESPONSE_TYPE": "<diagnosis | question>",
"RESPONSE_CONTENT": "<your diagnosis list or follow-up question>",
"CONFIDENCE": "<1-5>", // 5 = Very confident on the appropriateness of the response,
"RATIONALE": "<brief justification>"
}
```

Voting Prompt

```
You are a {voter['SPECIALIST']}. Another specialist ({candidate['SPECIALIST']}) proposed the
    following:

They suggested to proceed with a **{candidate['RESPONSE_TYPE']}**, which is:
"{candidate['RESPONSE_CONTENT']}"

Their rationale is:
"{candidate['RATIONALE']}"

Based on the current patient case, decide if you agree with the proposed next step.
Respond with ONLY one of the following options:
- AGREE
- DISAGREE
```

C Potential Risk

While DynamiCare is designed for research and educational purposes, there are potential risks associated with the misuse or misinterpretation of its outputs. The system simulates clinical reasoning using language models that, despite strong performance, may generate incorrect, incomplete, or misleading responses if taken at face value. If such models were applied in real-world medical settings without appropriate oversight, they could lead to harmful diagnostic decisions or delayed treatments. Additionally, since MIMIC-Patient is derived from real patient data (via the de-identified MIMIC-III database), there is a need to ensure responsible data handling and prevent re-identification risks, even though direct identifiers have been removed. Finally, overreliance on open-ended AI-generated diagnoses could inadvertently reinforce biases or amplify gaps in model training data. We emphasize that our framework is not intended for clinical deployment and should be used strictly in controlled, academic environments.