AI Hospital: Benchmarking Large Language Models in a Multi-agent Medical Interaction Simulator

Zhihao Fan[1], **Lai Wei**[2], **Jialong Tang**[1], **Wei Chen**[2*],
Siyuan Wang[3], **Zhongyu Wei**[4], **Jun Xie**[1], **Fei Huang**[1], **Jingren Zhou**[1]
[1]Alibaba Inc.
[2]School of Software Engineering, Huazhong University of Science and Technology
[3]University of Southern California
[4]School of Data Science, Fudan University
[1]fanzhihao.fzh@alibaba-inc.com, [*]lemuria_chen@hust.edu.cn

Abstract

Artificial intelligence has significantly revolutionized healthcare, particularly through large language models (LLMs) that demonstrate superior performance in static medical question answering benchmarks. However, evaluating the potential of LLMs for real-world clinical applications remains challenging due to the intricate nature of doctor-patient interactions. To address this, we introduce **AI Hospital**, a multi-agent framework emulating dynamic medical interactions between *Doctor* as player and NPCs including *Patient* and *Examiner*. This setup allows for more practical assessments of LLMs in simulated clinical scenarios. We develop the Multi-View Medical Evaluation (**MVME**) benchmark, utilizing high-quality Chinese medical records and multiple evaluation strategies to quantify the performance of LLM-driven *Doctor* agents on symptom collection, examination recommendations, and diagnoses. Additionally, a dispute resolution collaborative mechanism is proposed to enhance medical interaction capabilities through iterative discussions. Despite improvements, current LLMs (including GPT-4) still exhibit significant performance gaps in multi-turn interactive scenarios compared to non-interactive scenarios. Our findings highlight the need for further research to bridge these gaps and improve LLMs' clinical decision-making capabilities. Our data, code, and experimental results are all open-sourced at https://github.com/LibertFan/AI_Hospital.

1 Introduction

In recent years, large language models (LLMs) have achieved remarkable performance on medical question answering benchmarks (Jin et al., 2019; Gu et al., 2020; Pal et al., 2022; Chen et al., 2023b), rivaling even human experts (Singhal et al., 2022). However, significant challenges remain in applying LLMs to real-world clinical diagnosis. In practice, accurate diagnosis relies on multiple turns of interactions between doctors, patients, and medical staff. This typically involves initial patient consultations, followed by targeted medical examinations, and iterative information gathering to build a comprehensive clinical picture (Zhong et al., 2022; Chen et al., 2023e,d). This dynamic diagnostic process differs markedly from static medical Q&A datasets, where complete patient information is assumed to be available upfront. Despite the critical importance, research evaluating the performance of LLMs in these dynamic diagnostic scenarios remains scarce.

To explore the capabilities of LLMs in interactive clinical diagnosis, we introduce **AI Hospital**, an LLM-powered multi-agent framework designed to simulate real-world dynamic medical interactions. Following a minimalist design principle, AI Hospital consists of two non-player characters (NPCs), the *Patient* and the *Examiner*, along with one player character, represented by the *Doctor*. The point of interest is the LLM that plays the role of the *Doctor*, whose task is to complete the diagnosis of the *Patient* within a limited number of interaction turns and ultimately write a complete diagnosis report for the *Patient*. As starting with no prior knowledge of the *Patient*'s condition, to make accurate diagnosis, the *Doctor* need to efficiently ask medically insightful questions, and recommend correct medical examinations.

We further establish the **Multi-View Medical Evaluation (MVME)** benchmark based on AI Hospital framework, incorporating high-quality Chinese medical records screened by experienced professionals. The information in medical records is assigned to *Patient* and *Examiner*, and GPT-3.5 is leveraged to simulate their behaviors. The final diagnosis report generated by the *Doctor* will be compared with actual medical records to evaluate the *Doctor*'s interactive diagnostic ability. MVME

[*]Corresponding author (lemuria_chen@hust.edu.cn).

Proceedings of the 31st International Conference on Computational Linguistics, pages 10183–10213
January 19–24, 2025.

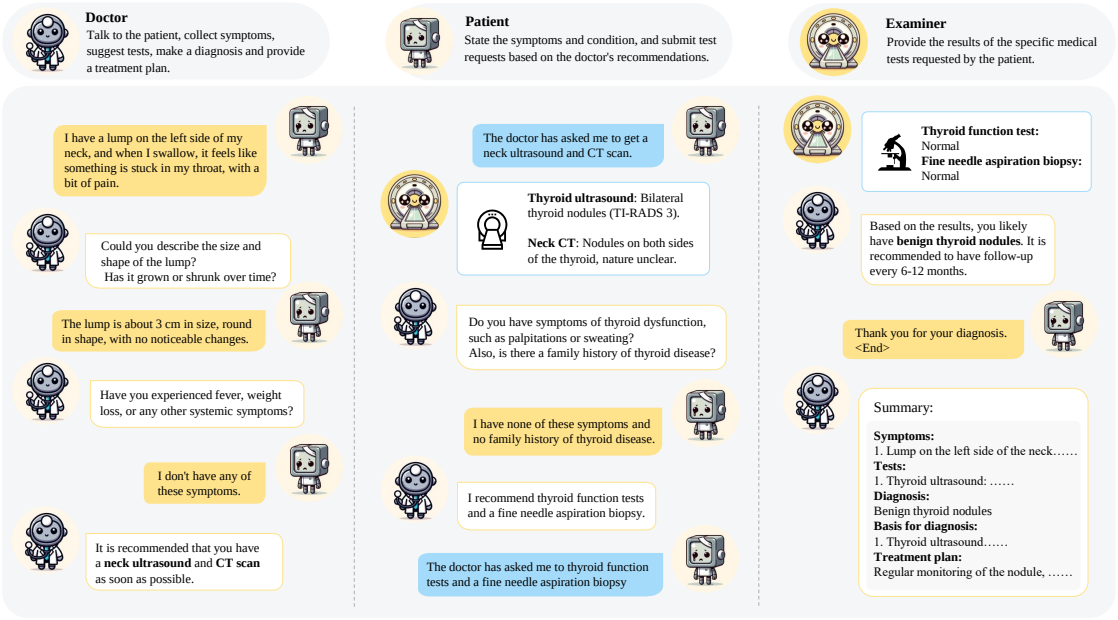

Figure 1: An example of multi-turn dialogue flow in AI Hospital, demonstrating the interactive process between the *Doctor* agent, *Patient* agent, and *Examiner* agent in a simulated clinical consultation.

benchmark evaluates the *Doctor*'s performance across three key dimensions: symptom collection, examination recommendation, and diagnosis and treatment. The evaluation methods include link-based automatic evaluation, model-based (GPT-4) evaluation, and human evaluation.

To enhance the diagnostic accuracy of LLMs, we further explore a collaborative mechanism inspired by previous research highlighting the importance of teamwork in clinical diagnosis (Croft et al., 2015; Centor et al., 2019; O'leary et al., 2010; Lamb et al., 2011). In this mechanism, multiple *Doctor*s independently diagnose the same case, generating distinct conversation trajectories and diagnostic reports. These *Doctor* agents then engage in discussions guided by a *Center Agent* to promote effective collaboration and reach consensus.

We conduct extensive experiments within AI Hospital, validating the feasibility and reliability of the framework and evaluating various LLMs in the interactive diagnostic process. Experimental results reveal a substantial performance gap between LLMs in interactive settings (where multiple rounds of interaction with *Patient*s are required) and one-step GPT-4 (which serves as an upper bound by accessing all information of *Patient* without interaction). In key metrics including diagnostic accuracy, reasoning, and treatment planning, the performance of GPT-4 and other LLMs in interactive settings falls **below 50%** of the performance achieved by GPT-4 in the one-step setting. The collaborative mechanism improves performance to a certain degree but still falls short of the upper bound. The main factors contributing to this suboptimal performance are the *Doctor* agent's failure to accurately recommend necessary medical examinations (thus lacking critical examination results) and the inability to inquire about key relevant symptoms during the interaction process. These findings highlight the challenges LLMs face in multi-turn medical diagnosis, particularly in gathering critical clinical information through interactions with *Patient*s.

In summary, the **main contributions** of this paper can be summarized as follows: 1) We introduce AI Hospital, to the best of our knowledge, **the first** LLM-powered multi-agent framework to simulate medical interactions, enabling comprehensive evaluation of LLMs' ability to navigate complex clinical scenarios; 2) We establish the Multi-View Medical Evaluation (MVME) benchmark, which leverages high-quality medical records to evaluate the performance of LLM-powered *Doctor* agents in collecting symptoms, recommending examinations, and making diagnoses; 3) We propose a dispute resolution collaborative mechanism that facilitates iterative discussions among *Doctor*s to enhance diagnostic accuracy. The potential broad applications of AI Hospital framework is comprehensively discussed in Appendix I.

2 Related Works

LLM Powered Agents Before the popularity of LLMs, there are already efforts to create agents in the medical field, particularly for medical education (Watts et al., 2021; Antel et al., 2022). However, these agents often lack flexibility, relying on rule-based or traditional machine learning algorithms made it difficult to accurately simulate the complexity of medical scenarios. The advancement of LLMs powered agents has led to significant strides in complex task resolution through human-like actions, such as tool-learning (Chen et al., 2023c; Schick et al., 2024), retrieval augmentation (Yue et al., 2023; Asai et al., 2023), role-playing (Park et al., 2023), communication (Xi et al., 2023; Wang et al., 2023c). This includes applications in software design and molecular dynamics simulation. Recent research (Tang et al., 2023; Wei et al., 2024) in the medical field has highlighted the critical roles and decision-making processes in medical QA, encompassing various investigations like CT scans, ultrasounds, electrocardiograms, and blood tests. Despite these advancements, effectively integrating LLM-based agents into the medical domain, particularly in disease diagnosis, presents a notable challenge (Zhou et al., 2023). Our research pioneers the use of multi-agent systems in creating a clinical diagnosis environment. We also introduce a novel mechanism for identifying, discussing, and resolving disputes in collaboration, demonstrating promising results in clinical diagnosis.

Large Language Models in Medicine Prior to the emergence of large language models (LLMs), the majority of automated diagnostic methods (Zhong et al., 2022; Chen et al., 2023e) relies on reinforcement learning to guide agents in gathering symptoms and conducting diagnoses. The development of LLMs in the medical domain has been driven by open-source Chinese LLMs and various fine-tuning methods. Models like MedPaLM (Singhal et al., 2022), DoctorGLM (Xiong et al., 2023), BenTsao (Wang et al., 2023a), ChatGLM-Med (Wang et al., 2023b), Bianque-2 (Chen et al., 2023f), ChatMed-Consult (Zhu and Wang, 2023), MedicalGPT (Xu, 2023), and DISC-MedLLM (Bao et al., 2023) are fine-tuned using different datasets, techniques, and frameworks, focusing on medical question answering, health inquiries and doctor-patient dialogues.

Evaluation in Medicine AI Prior research in medical AI evaluation has concentrated on non-interactive tasks, including question answering, entity and relation extraction, and medical summarization and generation. In biomedical question answering, key datasets such as MedQA (USMLE) (Jin et al., 2021), PubMedQA (Jin et al., 2019), and MedMCQA (Pal et al., 2022) are utilized, with accuracy serving as the primary evaluation metric. The objective of entity and relation extraction (Li et al., 2020) is to categorize named entities and their relationships from unstructured text into specific predefined classes. Prominent biomedical NER datasets include NCBI Disease (Doğan et al., 2014), JNLPBA (Collier and Kim, 2004), BC5CDR (Li et al., 2016), BioRED (Luo et al., 2022) and IMCS-21 (Chen et al., 2023b,d), with the F1 score being the standard for model performance assessment. Medical summarization and generation tasks involve converting structured data, like tables, into descriptive text. This includes the creation of patient clinic letters, radiology reports, and medical notes (Liu et al., 2023). The principal datasets for these tasks are PubMed (Jin et al., 2019) and MentSum (Sotudeh et al., 2022). A recent study introduced BioLeaflets (Yermakov et al., 2021) and assessed multiple Large Language Models (LLMs) in data-to-text generation.

3 Setup of AI Hospital

As depicted in Figure 1, the AI Hospital framework comprises two **NPC characters** — the *Patient*, the *Examiner* — and one **player character**, the *Doctor*. Each character assumes specific roles and responsibilities within the framework. The *Doctor* communicates with the *Patient* to gather symptoms, medical history, etc., suggests necessary medical tests, and ultimately provides a diagnosis and treatment plan based on the collected information. The *Patient* responses to *Doctor*'s questions and, upon receiving test recommendations, requests the specified medical tests from the *Examiner*. The *Examiner*, in turn, provides the results of the requested medical tests to the *Patient*. The interaction between the agents is limited to a predetermined maximum number of turns, set to **10 turns** in this paper. If the *Doctor* reaches a diagnosis before 10 turns, the conversation will conclude earlier.

3.1 Agents Setup with Medical Records

Medical records are valuable resource for reconstructing the hospital visit experience and simulat-

ing real-world medical interactions. By leveraging these medical records, we can reverse-engineer the diagnostic process and shape the behavior of agents within the AI Hospital framework. We categorize the information in each medical record into three types: 1) **Subjective Information** This category includes the patient's symptoms, etiology, past medical history, habits, etc., which are primarily provided by the patient during their verbal interactions with the doctor; 2) **Objective Information** This category encompasses the results of medical tests such as *Complete Blood Counts*, *Urinalysis*, and *Chest X-rays*. The presence of these data in medical records indicates that doctors recommended these medical tests to patients during the diagnostic process; 3) **Diagnosis and Treatment** This category consists of *diagnostic results*, *diagnostic rationales*, and *treatment courses*, which are the final diagnostic reports made by the doctor during the diagnostic process, based on the combination of **subjective** and **objective** information.

The AI Hospital framework assigns information from medical records to each agent in a manner that aligns with real-world scenarios. In a typical hospital setting, patients are only aware of their subjective experiences and must rely on doctors to order and interpret medical tests. Thus, the *Patient* agent in the AI Hospital framework is set to has access only to subjective information, and the *Examiner* agent holds the objective information, representing the healthcare professionals who perform the medical tests. The *Doctor* agent starts without any information, reflecting the fact that doctors must gather any relevant information through interactions with patients. This distribution of information among the agents mirrors the real-world flow of information in a medical diagnostic process, ensuring a realistic simulation.

3.2 Agent Behavior Setting for NPCs

In the AI Hospital framework, we leverage GPT-3.5 to power *Patient* and *Examiner* agent, enabling them to embody their roles authentically. Beyond providing NPCs with corresponding information in medical records, we also employ meticulous prompt engineering to encourage they exhibit more realistic behavior patterns.

Patient The *Patient* agent may not proactively disclose relevant physical conditions, but they will provide truthful responses when the doctor asks specific questions. If the *Doctor* recommends a specific medical examination, the agent will comply and undergo the suggested examination. The agent may use colloquial language. The prompts for the *Patient* agent is shown in Table 15.

Examiner The *Examiner* agent's primary task is to provide relevant examination results when the *Patient* agent requests a query for a specific medical test. Upon receiving an examination query, the agent first identifies the requested medical examination and rejects any request that is ambiguous or unclear or does not specify the examination name. If the corresponding medical examination results are available, the *Examiner* agent returns the relevant findings to the doctor. In cases where no specific results are found, the agent reports no abnormalities. The prompts for the *Examiner* agent are shown in Table 16 and 18.

3.3 Agent Behavior Setting for Player

The player agent, i.e., the *Doctor*, can be powered by various LLMs that are being evaluated. However, in order to be able to engage in conversations based on predefined settings, LLMs are required to be well instruction-followed, otherwise LLMs will struggle to interact in AI Hospital.

Doctor The *Doctor* agent is encouraged to actively gather information, focusing on obtaining the patient's physical conditions like symptoms and medical history. A crucial aspect of the agent's role is to recommend necessary medical examinations when the agent believes that additional objective information are necessary to make a confident diagnosis or to confirm a suspected condition. By synthesizing both subjective and objective findings, the agent aims to make correct diagnose, mirroring the systematic decision-making process employed by experienced doctors. The prompts for the *Doctor* agent is shown in Table 22.

3.4 Dialogue Flow in AI Hospital

The AI Hospital framework simulates a realistic diagnostic process through a structured dialogue flow involving multiple agents. The conversation is initiated by the *Patient* agent, who first presents a chief complaint, which only contains a small part of the subjective information. Notably, we specify that while the patient actively provides the chief complaint to the doctor initially, such proactive behavior is not guaranteed thereafter. The *Doctor* agent then engages in a series of interactions with the *Patient* and *Examiner* agents to gather necessary information and make an accurate diagnosis.

Throughout the dialogue, each agent's responses are prefixed with special symbols to explicitly indicate the intended recipient of their message, enabling a seamless multi-party conversation flow. For a more detailed description of the dialogue flow, please refer to Appendix B.

4 MVME: Evaluation of LLMs as Doctors for Clinical Diagnosis

Based on AI Hospital, we assess the feasibility of employing various LLMs as *Doctor* agent for clinical diagnosis by establishing the Multi-View Medical Evaluation (MVME) benchmark.

4.1 Multi-View Evaluation Criteria

After the diagnosis process is completed, the *Doctor* agent is required to generate a diagnostic report for the *Patient* based on the entire conversation trajectory. It's required that the diagnostic report consists of 5 parts, including the patient's *Symptoms*, *Medical Examinations*, *Diagnostic Results*, *Diagnostic Rationales*, and *Treatment Plan*.

Link-based Evaluation We compute entity-overlap-based automated metrics for the *Diagnostic Results* section. We extract all disease entities from the diagnostic results provided by the *Doctor* agents and the actual medical records, and link them to their corresponding standardized disease entities by International Classification of Diseases (ICD-10) (Trott, 1977). We then calculate the entity overlap to measure the accuracy of the final diagnoses made by *Doctor* agents. We report the average number of extracted disease entities (#), set-level precision (P), recall (R), and F1 score (F) metrics.

Model-based Evaluation In addition to the above link-based evaluation method, we also utilize GPT-4 as the model-based evaluator to compare each part of the diagnostic report generated by the *Doctor* agent with the raw medical record, using a discrete scoring system from 1 to 4 (poor to excellent). With carefully designed evaluation prompts, we establish specific scoring criteria for each section: the evaluation for the *Symptoms* section is designed to reflect the completeness of symptom collection during interactions, the *Physical Examination Results* section is crafted to evaluate the accuracy of recommended examinations, and other sections are established to measure the comprehensiveness of *Doctor*'s diagnosis and treatment. These metrics well reflect the dynamic medical decision-making capabilities, encompassing proactive questioning, information gathering, clinical knowledge, and comprehensive judgment.

Human Evaluation To validate the reliability of GPT-4 based evaluator, we also employ parallel human evaluation with the help of professional physicians who follow the identical scoring criteria provided in the prompts of GPT-4 evaluator.

Specialty Department	No. of Cases (%)
Surgery	180 (35.6%)
Internal Medicine	153 (30.2%)
Obstetrics and Gynecology	94 (18.6%)
Pediatrics	29 (5.7%)
Otorhinolaryngology	23 (4.5%)
Others	27 (5.3%)

Table 1: Distribution of case records across specialty departments.

4.2 MVME Dataset Construction

We collect diverse medical records from various departments on iiyi.com, a website that compiles an extensive database of clinical cases in Chinese. Each case can be divided into three main components: **subjective information** (including history of present illness, personal history, and past medical history), **objective information** (including physical examination and auxiliary examinations), and the **doctor's diagnosis and treatment process** (covering diagnostic results, diagnostic rationales, and treatment plan). After eliminating records with incomplete information, a total of 506 high-quality case records remained. Table 1 shows the detailed distribution of these cases across different medical specialty departments.

4.3 Dataset Visualization and Statistics

The dataset encompasses a diverse range of medical specialties, subspecialties, diseases, examinations, and symptoms. It covers **12** specialties, **48** subspecialties, and a wide variety of diseases, with *Type 2 Diabetes Mellitus*, *Arrhythmia*, *Hypertension*, and *Hyperlipidemia* being among the most prevalent (Figure 3). On average, each patient in the dataset undergoes **3.5** medical examinations, with a total of **769** unique examination items covering various types of tests (Figure 4, left subfigure). The dataset also features an average of **6.8** symptoms per case, with patients' chief complaints including an average of **1.7** symptoms. The symptoms span multiple

body systems, with over **960** unique symptoms represented in the dataset (Figure 4, right subfigure). These statistics underscore both the comprehensive coverage of our dataset and the substantial challenge it presents - *Doctor* agents must not only identify relevant medical examinations from hundreds of options but also effectively elicit additional symptoms beyond the initial complaints through strategic questioning. More detailed analysis of the dataset can be found in Appendix A.

4.4 Dataset Quality Assessment

To validate the quality of the collected medical records, we select samples from the 10 most common subspecialty departments, randomly choosing 5 cases per department for review, which accounts for nearly half of the total sample size. Doctors from the corresponding departments are hired to evaluate the "Diagnosis and Treatment" section, including the diagnostic result, diagnostic rationale, and treatment plan. They are asked to make a binary choice, classifying each section as either "fundamentally correct" or "obviously incorrect". If all three parts of a medical record are deemed fundamentally accurate, then the medical record is considered correct. The expert validation process concludes that **94%** of the reviewed records are correct, indicating a high level of accuracy and reliability in the collected data. More details can be found in Appendix E.

5 Collaborative Diagnosis of LLMs Focused on Dispute Resolution

To further improve diagnostic accuracy, we propose a collaborative mechanism for clinical diagnosis that leverages the power of multiple LLMs. In our collaborative framework, we employ different LLMs to serve as individual *Doctor*s, each engaging in independent interactive consultations with the *Patient*, resulting in diverse dialogue trajectories and diagnostic reports. To streamline the process of forming a unified diagnostic report, we introduce a *Central Agent* to participate as a moderator. We provide a detailed description of the collaborative mechanism in Appendix C.

The *Central Agent* consolidates and analyzes the data collected from multiple *Doctor*s, confirms disputed points with *Patient* and *Examiner*, and synthesizes a comprehensive summary of the patient's condition. Through multiple discussion iterations, the *Central Agent* identifies key points of disagreement among *Doctor*s and guides them to engage in targeted discussions, progressively refining their understanding and working towards a consensus. This collaborative mechanism harnesses the collective intelligence of LLMs to enhance the accuracy and robustness of clinical diagnosis by capitalizing on their diverse knowledge and reasoning capabilities while promoting a structured and iterative process of refining diagnostic reports. The entire process is described in pseudocode form in Algorithm 1, and the prompts designed for the *Central Agent* and *Doctor*s during the collaborative process are listed in Table 21 and Table 23 in the appendix.

6 Experiments

6.1 Agent Behavior Analysis in AI Hospital Framework

Before presenting the main results, it is crucial to verify whether the agents in the AI Hospital framework effectively align with their intended roles and behaviors. We conduct a experiment to investigate the behaviors of several key agents, including the *Patient*, *Examiner*, and *Doctor*.

Evaluation Metric For the Patient agent, we focus on two dimensions in the communication between the *Patient* and the *Doctor*. The first dimension is the **relevance** of the *Patient*'s responses to the *Doctor*'s questions. The second dimension is the **honesty** of the *Patient*'s responses with the subjective information in the medical record. For the *Examiner* agent, we assess the **accuracy** of the agent's understanding of the requested medical examination and its ability to return the corresponding examination results when receiving a query for a medical examination. For the *Doctor* agent, we evaluate the **consistency** of the final diagnostic report with the information in the dialogue flow. We categorize the consistency into three levels: 1) significantly inconsistent, 2) slightly inconsistent, and 3) mostly consistent. These levels are assigned scores of 1, 2, and 3, respectively. Finally, we map this score to a range of 0-100. We document our evaluation methodology in detail in Appendix G.

Experimental Setup We employ multiple *Doctor* agents, including GPT-3.5 and GPT-4 (OpenAI, 2023), Wenxin-4 (Baidu, 2023), and Qwen-Max (Bai et al., 2023). We randomly select 50 medical record samples and ask each agent generate 50 multi-turn dialogue trajectories within the AI Hospital framework. We manually label all the metrics and report the average values.

	Patient			Examiner		Doctor
	#	Relevance	Honesty	#	Accuracy	Consistency
Qwen-Max	429	100.0%	99.0%	56	98.2%	99.0
Wenxin-4	472	100.0%	98.1%	68	98.5%	99.0
GPT-3.5	417	100.0%	99.5%	57	98.2%	98.0
GPT-4	378	100.0%	99.7%	61	100.0%	100.0

Table 2: Human evaluation for agent behavior in AI Hospital. # represents the sample size, such as number of total doctor-patient QA pairs in 50 dialogues.

	Symptoms	Medical Examinations	Diagnostic Results	Diagnostic Rationales	Treatment Plan
	Interaction				
Baichuan-13B (Yang et al., 2023)	52.56 (2.77)	22.07 (2.57)	19.50 (2.74)	17.40 (2.51)	13.97 (2.37)
HuatuoGPT-II-13B (Chen et al., 2023a)	61.06 (2.17)	29.37 (2.30)	20.03 (2.56)	20.03 (2.37)	14.23 (2.18)
HuatuoGPT-II-34B (Chen et al., 2023a)	68.43 (1.83)	32.40 (2.37)	25.20 (2.52)	27.46 (2.55)	21.33 (2.37)
GPT-3.5 (OpenAI, 2023)	66.39 (1.33)	40.63 (2.57)	23.90 (2.43)	24.43 (2.42)	17.73 (2.17)
Wenxin-4 (Baidu, 2023)	67.79 (1.33)	33.93 (2.50)	26.23 (2.63)	26.46 (2.57)	22.00 (2.43)
Qwen-Max (Bai et al., 2023)	61.69 (2.10)	35.50 (2.63)	26.46 (2.63)	28.76 (2.63)	24.90 (2.45)
GPT-4 (OpenAI, 2023)	**69.03** (1.27)	**40.83** (2.30)	**29.36** (2.58)	**30.76** (2.57)	**26.93** (2.63)
	Collaboration				
2 Doctors w/o DR (GPT-3.5+GPT-4)	75.49 (2.03)	43.03 (3.03)	35.56 (2.83)	38.53 (2.76)	32.40 (2.60)
2 Doctors (GPT-3.5+GPT-4)	78.06 (1.83)	43.60 (2.77)	38.06 (2.72)	41.56 (2.75)	35.90 (2.62)
3 Doctors (GPT-3.5+GPT-4+Wenxin-4)	**80.26** (1.80)	**52.70** (2.83)	**39.60** (2.80)	**44.23** (2.77)	**37.26** (2.63)
	One-Step				
GPT-4*	100.0*	100.0*	58.89 (1.63)	66.59 (1.33)	53.16 (1.83)

Table 3: MVME: GPT-4 evaluation with reference in clinical consultation. GPT-4* in One-Step is the upper bound. For GPT-4*, the ground truth of symptoms and medical examinations are provided, resulting in a score of 100.0.

Results and Analysis Table 2 demonstrates the effectiveness of the AI Hospital framework in simulating realistic medical interactions, with high scores (**all over 95**) across all metrics indicating reliable and consistent agent behaviors. The *Patient* agent can provide accurate and pertinent information, the *Examiner* agent can accurately understand and return requested medical examination results, and the *Doctor* agent can generate consistent diagnostic reports. Above results validate the reliability and effectiveness of the proposed multi-agent system, laying a solid foundation for assessing LLMs' performance in clinical diagnosis.

6.2 Performance of Various *Doctor* Agents

Based on the AI Hospital, We evaluate a range of LLMs as *Doctor* agents, including GPT (OpenAI, 2023) (GPT-3.5 and GPT-4), Wenxin-4 (Baidu, 2023), QWen-Max (Bai et al., 2023), Baichuan 13B (Yang et al., 2023), HuatuoGPT-II 13B and 34B (Chen et al., 2023a). Among these, HuatuoGPT-II represents the leading medical LLMs, passing multiple Chinese medical licensing exams and outperforming GPT-4 in various Chinese medical scenarios. We specifically chose HuatuoGPT-II for comparison as most other medical LLMs show limited instruction-following capabilities during training, making them unsuitable for customized prompts and effective dialogue in our benchmark testing.

Evaluation As mentioned in § 4.1, we employ the proposed multi-view evaluation criteria. We normalize the scores of all metrics to a range between 0 and 100 and utilize the classic bootstrap method (Efron, 1992) to compute the variance.

One-Step Diagnosis as Upper Bound In the one-step diagnosis, we directly feed the patient's subjective information and objective information described in § 3.1 as input to GPT-4, prompting it to generate a diagnostic report without going through the interactive diagnostic phase. We consider the performance of GPT-4 in this one-step setting as the upper bound of what LLMs can achieve in scenarios requiring interaction.

Interactive Diagnostic Performance The main experimental results are presented in Table 3 and Table 4. One of notable observations is that the diagnostic performance of existing LLMs in

	#	R	P	F1
	Interaction			
Baichuan (13B)	1.58	10.21	23.79	14.28
HuatuoGPT-II (13B)	1.72	12.76	24.84	16.85
HuatuoGPT-II (34B)	1.86	17.48	30.95	22.34
GPT-3.5	1.81	19.19	37.39	25.37
Wenxin-4	2.50	22.03	31.44	25.91
Qwen-Max	1.77	**22.42**	43.38	29.56
GPT-4	1.52	21.64	**50.26**	**30.26**
	Collaboration			
2 Doctors w/o DR	2.37	28.44	41.45	33.74
2 Doctors	2.41	29.51	**43.62**	35.21
3 Doctors	3.20	**36.54**	39.58	**38.00**
	One-Step			
GPT-4*	2.30	38.90	58.97	46.88

Table 4: MVME: Link-based evaluation of diagnostic results.

the AI Hospital framework falls significantly short of the upper bound set by the one-step GPT-4 approach. Even GPT-4 achieves less than **50%** of the upper bound performance. This finding highlights the substantial limitations of current LLMs in interactive settings, suggesting that they have not yet learned sufficiently rich real-world clinical decision-making experiences. We also observe that LLMs with less parameters tend to exhibit weaker interactive abilities, such as Baichuan-13B and HuatuoGPT-II-13B, demonstrates lower performance in interactive diagnosis.

Analysis of Factors Affecting Diagnostic Performance Based on Table 3, we further explore the relationship between the information finally collected and the quality of diagnosis. We use Symptoms and Medical Examinations to measure the completeness of patient information, and use Diagnostic Results, Diagnostic Rationales, and Treatment Plans to evaluate diagnostic quality. By fitting a simple linear regression, we present our results in Figure 7 (Appendix D.1), which show that there is a significant positive correlation between more complete patient information and higher diagnostic quality. This further explains the shortcomings of current LLMs, that is, it is difficult for LLMs to collect patients' symptoms through active questioning, and it is even more difficult for them to recommend correct medical examinations. This lack of dynamic clinical decision-making ability is a huge obstacle that prevents LLMs from diagnosing like real doctors.

Performance Across Departments Our analysis of the performance of various LLMs across different hospital departments reveals that the positive correlation between interaction ability and diagnostic ability is more prominent when considering larger scale variations (Table 6, 7, 8). The overall performance of LLMs varies across different departments, with most models performing better in the SURG and ENT departments compared to others, particularly the PEDS department.

Human Evaluation To evaluate the effectiveness of the model-based evaluation using GPT-4, we compare its results with human evaluation on 50 randomly selected summary reports. The human evaluation follows the same scoring system used in the model-based evaluation. The results of the human evaluation are very close to those of the GPT-4 evaluation across the five different aspects, with differences of less than or equal to 4%, indicating that GPT-4 is capable of demonstrating performance comparable to human evaluation (Figure 8).

Other LLMs as Evaluator To eliminate the potential preference of GPT-4 evaluations for outputs generated by GPT-4, we also include Qwen and Deepseek as additional evaluators (Tables 10 and 11). We find that the results of using Qwen-Max as an evaluator tend to award higher scores to outputs generated by Qwen-Max. For Deepseek, which may be fairer since our baseline does not include Deepseek, we find that its scoring is relatively closer to the results presented in Table 2. More detailed analysis of the evaluation can be found in Appendix D.

7 Further Analysis

7.1 Collaboration Mechanism

In Table 3, we also evaluate several methods with different settings of the cooperation mechanism. The comparative methods include collaborative diagnosis with 3 and 2 Agents, an 2 Agents without Dispute Resolution. They are denoted as 3 Doctors, 2 Doctors and 2 Doctors w/o DR. The initial two *Doctor* agents are powered by GPT-3.5 and GPT-4 for interactive consultation, while the last agent uses Wenxin-4.

Effectiveness Collaboration Mechanism We observe several key findings: 1) The collaborative use of LLMs can exceed the performance of single GPT-4, thereby validating the efficacy of the cooperative mechanism; 2) Collaboration among "3 Doctors" enhances diagnosis compared to "2 Doc-

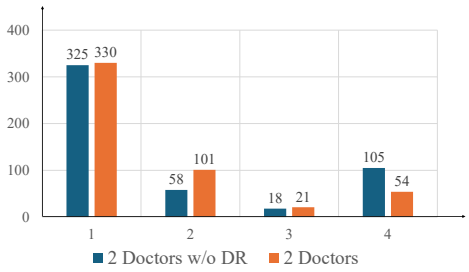

Figure 2: Statistical analysis of discussion rounds in collaborative frameworks with and without "Dispute Resolution" mechanism.

Error Type	#
Omission of Necessary Medical Examinations	99
Ignoring Potential Symptom Associations	52
Erroneous Judgment	68

Table 5: Classification and statistics of misdiagnoses (1 point) of the *Doctor* agent powered by GPT-4.

tors", highlighting the benefits of more agents in cooperation; 3) The removal of the "Dispute Resolution" mechanism from the "2 Doctors" reduces its effectiveness, emphasizing the significance of establishing a better consensus.

Efficiency of Dispute Resolution in Collaboration For the "Dispute Resolution", we continue to check whether the *Doctor* agents can reach consensus more rapidly. In terms of efficiency, a comparative analysis is conducted on the number of discussion rounds necessary to achieve consensus, both with and without the "Dispute Resolution" mechanism. The outcomes are detailed in Figure 2. These findings reveal a marked increase in the rate of consensus achieved within the initial four discussion rounds following the adoption of the dispute resolution mechanism. This enhancement suggests that the process, facilitated by the *Central Agent* highlighting controversial issues and multiple *Doctor* agents concentrating on these discussions, effectively reduces the time required to achieve consensus.

7.2 Reasons for Failure Cases

We analyze an analysis on 219 cases where GPT-4 render incorrect diagnostic results, and rated as 1 point by GPT-4. Through a systematic manual review (performed by human professional doctors), these errors are mainly categorized into three distinct types, which are detailed in Table 5.

Omission of Necessary Medical Examinations An illustrative case involved the failure to detect gallbladder stones, attributed to the absence of a recommended abdominal ultrasound. This category highlights instances where GPT-4 did not suggest essential medical examinations that could have potentially confirmed or ruled out possible medical conditions.

Ignoring Potential Symptom Associations In certain cases, GPT-4 focuses only the symptoms given by the patient, such as soft tissue swelling in the feet, while ignoring underlying complications, such as diabetes. This type of error arises from the LLMs' limited recognition of the interconnectedness between symptoms and underlying health issues, and its failure to prompt further inquiry into the patient's comprehensive health status.

Erroneous Judgment Even when presented with relatively complete symptomatology and medical examination results, GPT-4 occasionally reach incorrect conclusions. This category of error points to a lack of sufficient medical expertise embedded within the LLMs, leading to diagnostic inaccuracies even with comprehensive data.

8 Conclusion

In this paper, we focus on quantifying LLMs' capabilities in interactive clinical diagnosis, in contrast to traditional static Medical QA datasets. The AI Hospital framework (*Patient* and *Examiner* acting as NPC agents and *Doctor* acting as player agent) and MVME benchmark (506 real-world complete medical records) are introduced, which simulates medical interactions well, shown by reliable agent behaviors. Evaluating LLM-powered doctor agents reveals key points. Their performance in the AI Hospital's interactive setting lags far behind one-step GPT-4, highlighting current LLMs' limitations in dynamic clinical decision-making. LLMs with fewer parameters have weaker medical interactive ability. Also, gathering complete patient info correlates with better diagnosis, yet current LLMs struggle with this dynamic information-gathering process. Further evaluation shows LLM performance also varies across different medical departments. GPT-4's evaluation is comparable to human evaluation, and using other LLMs as evaluators uncovers biases. In summary, LLMs have made progress but face challenges in clinical diagnosis. Future efforts should focus on developing better training for LLMs to close the gap with human experts in clinical medicine.

Limitations

This study has several important limitations that should be considered. Firstly, the dataset is mainly sourced from Chinese medical records, which may limit the generalizability of the findings to other languages and medical systems. Secondly, the impact of various patient agent settings like different patient backgrounds, cultures, and biases on model performance remains unexamined. Thirdly, the study doesn't explore the doctor agents' ability to utilize external tools, external knowledge, or make decisions based on multimodal medical information. Moreover, relying on numerous LLM APIs for testing new models consumes a large amount of resources and potentially increases carbon emissions. Finally, the AI Hospital and collaborative mechanism proposed is based on a relatively simple framework and might not fully capture the complexity of real-world clinical collaboration scenarios, requiring further refinement and validation in more diverse and practical settings.

Ethics Consideration

We recognize the potential implications of our work and have taken steps to address them. Firstly, to ensure transparency and reproducibility, we have released the publicly accessible online medical records data used in our study. The data sources have undergone a process of de-identification, removing sensitive information before our collection. Furthermore, we recognize the potential for bias in AI systems, which could perpetuate or amplify disparities in healthcare. To mitigate this risk, we have made efforts to ensure the diversity and representativeness of our medical record datasets. . By proactively addressing these considerations, we aim to realize the potential benefits of AI-assisted diagnosis while ensuring its responsible and equitable implementation.

Acknowledgments

This research was supported by *National Natural Science Foundation of China* (No.62406121) and *Natural Science Foundation of Hubei Province, China* (No.2024AFB189).

References

Ryan Antel, Samira Abbasgholizadeh-Rahimi, Elena Guadagno, Jason M Harley, and Dan Poenaru. 2022. The use of artificial intelligence and virtual reality in doctor-patient risk communication: A scoping review. *Patient Education and Counseling*, 105(10):3038–3050.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baidu. 2023. wenxin4.0. https://wenxin.baidu.com/.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *Preprint*, arXiv:2308.14346.

Robert M Centor, Rabih Geha, and Reza Manesh. 2019. The pursuit of diagnostic excellence. *JAMA network open*, 2(12):e1918040–e1918040.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, et al. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms. *arXiv preprint arXiv:2311.09774*.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023b. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. 2023c. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Wei Chen, Shiqi Wei, Zhongyu Wei, and Xuan-Jing Huang. 2023d. Knse: A knowledge-aware natural language inference framework for dialogue symptom status recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10278–10286.

Wei Chen, Cheng Zhong, Jiajie Peng, and Zhongyu Wei. 2023e. Dxformer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations. *Bioinformatics*, 39(1):btac744.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Huimin Zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023f. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *Github*.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78.

Peter Croft, Douglas G Altman, Jonathan J Deeks, Kate M Dunn, Alastair D Hay, Harry Hemingway, Linda LeResche, George Peat, Pablo Perel, Steffen E Petersen, et al. 2015. The science of clinical practice: disease diagnosis or patient prognosis? evidence about "what is likely to happen" should shape clinical practice. *BMC medicine*, 13(1):1–8.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Bradley Efron. 1992. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer.

Qizheng Gu, Cong Nie, Ruixiang Zou, Wei Chen, Chaojun Zheng, Dongqing Zhu, Xiaojun Mao, Zhongyu Wei, and Dong Tian. 2020. Automatic generation of electromyogram diagnosis report. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1645–1650. IEEE.

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*.

Benjamin W Lamb, Helen WL Wong, Charles Vincent, James SA Green, and Nick Sevdalis. 2011. Teamwork and team performance in multidisciplinary cancer teams: development and evaluation of an observational assessment tool. *BMJ quality & safety*, 20(10):849–856.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Jialin Liu, Changyu Wang, and Siru Liu. 2023. Utility of chatgpt in clinical practice. *Journal of Medical Internet Research*, 25:e48568.

Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. 2022. Biored: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23(5):bbac282.

KJ O'leary, CD Ritter, H Wheeler, MK Szekendi, TS Brinton, and MV Williams. 2010. Teamwork on inpatient medical units: assessing attitudes and barriers. *BMJ Quality & Safety*, 19(2):117–121.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. Mentsum: A resource for exploring summarization of mental health online posts. *arXiv preprint arXiv:2206.00856*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, 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*.

PA Trott. 1977. International classification of diseases for oncology. *Journal of clinical pathology*, 30(8):782.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Chatglm-med. https://github.com/SCIR-HI/Med-ChatGLM.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023c. A survey on large

language model based autonomous agents. *arXiv preprint arXiv:2308.11432*.

Penni I Watts, Donna S McDermott, Guillaume Alinier, Matthew Charnetski, Jocelyn Ludlow, Elizabeth Horsley, Colleen Meakim, and Pooja A Nawathe. 2021. Healthcare simulation standards of best practicetm simulation design. *Clinical Simulation in Nursing*, 58:14–21.

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2024. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Ruslan Yermakov, Nicholas Drago, and Angelo Ziletti. 2021. Biomedical data-to-text generation via fine-tuning transformers. *arXiv preprint arXiv:2109.01518*.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Wei Lin, et al. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *arXiv preprint arXiv:2309.11325*.

Cheng Zhong, Kangenbei Liao, Wei Chen, Qianlong Liu, Baolin Peng, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. Hierarchical reinforcement learning for automatic disease diagnosis. *Bioinformatics*, 38(16):3995–4001.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Wei Zhu and Xiaoling Wang. 2023. Chatmed: A chinese medical large language model. https://github.com/michael-wzhu/ChatMed.

A Visual Analysis Details of the Dataset

To demonstrate the richness and diversity of our dataset, we conduct a comprehensive analysis focusing on the distribution of specialties, subspecialties, diseases, medical examinations, and symptoms.

Figure 3 presents a pie chart of the medical specialties (left subfigure) and a wordcloud of the diseases (right subfigure) in our dataset. Figure 5 further illustrates the distribution of medical subspecialties using a histogram. Our dataset encompasses a total of 12 specialties (e.g., *Surgery*, *Internal Medicine*, and *Obstetrics and Gynecology*) and 48 subspecialties (e.g., *Orthopedics* and *urology* under *Surgery*; *Gastroenterology* and *Neurology* under *Internal Medicine*). The overall distribution exhibits a long-tail pattern. The wordcloud of diseases reveals that *Type 2 Diabetes Mellitus*, *Arrhythmia*, *Hypertension*, and *Hyperlipidemia* are among the most prevalent diseases. As each patient may have multiple comorbidities, the medical records are highly heterogeneous, with almost no identical cases in terms of disease composition. This showcases the diversity of our dataset.

Figure 4 displays the distribution of medical examinations (left subfigure) and symptoms (right subfigure) in our dataset. On average, each patient in the dataset takes **3.5** medical examinations, with every patient having at least **1** medical examination record. This highlights the complexity of our cases, as they are not simple instances of common cold or fever that can be easily diagnosed based on a few symptoms. Our dataset comprises *769* unique medical examination items, including hematological tests, radiological imaging, functional assessments, and histopathological examinations. Regarding symptoms, each case contains an average of **6.8** symptoms, with the patient's chief complaint including an average of **1.7** symptoms. This finding is consistent with previous research (Wei et al., 2018; Chen et al., 2023e), indicating that patients struggle to express all their symptoms at one time, requiring doctors to gather new symptoms through interactive processes. The symptoms cover various systems, including gastrointestinal, respiratory, cardiovascular, neurological, genitourinary, and musculoskeletal, encompassing over *960* unique symptoms.

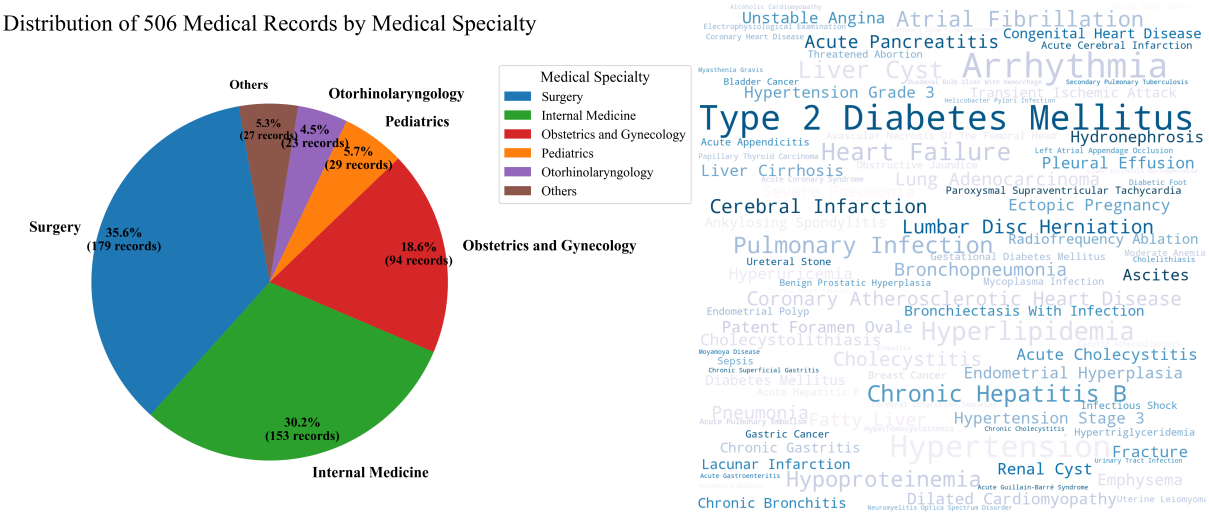

Figure 3: Distribution of specialties and diseases in the dataset. The left subfigure displays a pie chart showing the proportion of cases across top 5 most common specialties, while the right subfigure presents a word cloud illustrating the prevalence of diseases, with larger font sizes indicating higher frequencies.

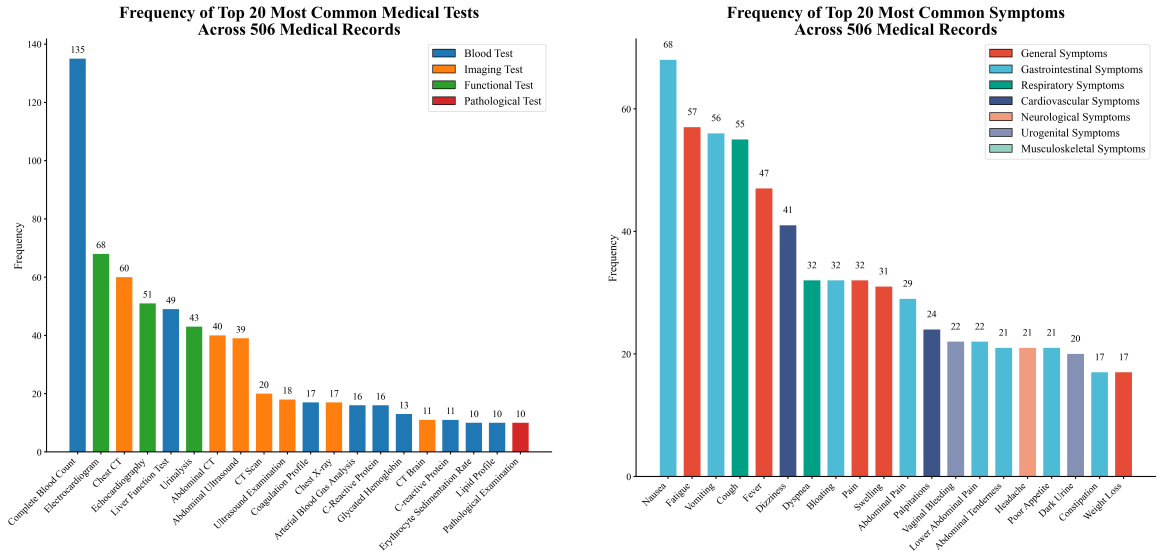

Figure 4: Distribution of medical examinations and symptoms in the dataset. The left subfigure displays a histogram showing the frequency of unique medical examination items, totaling 769, while the right subfigure presents a histogram depicting the frequency of unique symptoms, encompassing over 960 symptoms across various body systems.

It is important to note that the aforementioned values, such as 769 unique medical examination items and over 960 unique symptoms, may be overestimated. Due to the lack of readily available standardized tools, we rely entirely on GPT-4 to extract and standardize the names of medical examinations and symptoms from raw medical records. Consequently, some medical examinations or symptoms may not be fully standardized, leading to higher statistical results. However, the average statistical results should be reasonably accurate, reflecting the average number of unique symptoms and medical examinations per case. This highlights the challenges faced by physician agents when interacting with patient agents, as they need to gather sufficient information to make accurate diagnostic decisions.

B Detailed Description of the Dialogue Flow in AI Hospital

This section is the detailed description of the dialogue flow in AI Hospital in §3.4. The AI Hospital framework aims to simulate a realistic diagnostic

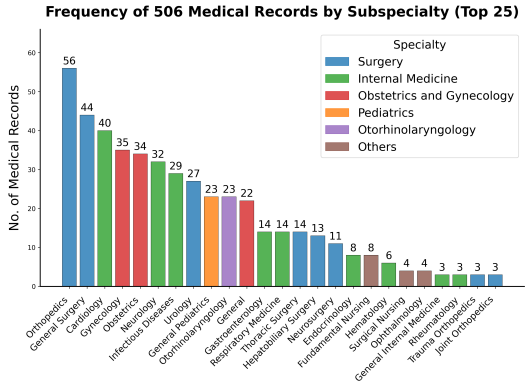

Figure 5: Histogram presenting the distribution of cases across top 25 most common subspecialties, revealing a long-tail pattern in the dataset.

process by orchestrating a structured dialogue flow involving multiple agents, namely the *Doctor*, *Patient*, and *Examiner*. This appendix provides a comprehensive description of the dialogue flow, detailing the initiation, progression, and termination phases.

Dialogue Initiation The dialogue commences with the *Patient* agent generating a chief complaint based on the information contained in their medical record. To create this initial complaint, GPT-4 is employed to analyze the patient's medical record and generate a concise statement that encapsulates the patient's recent physical condition. The generated complaint is designed to align with the predefined persona of the patient, accurately reflecting their language style and focusing on a relevant subset of the subjective information available in the medical record. This chief complaint serves as the starting point for the first round of dialogue between the *Patient* and *Doctor* agents.

Dialogue Progression The diagnostic process unfolds through a series of interactions between the *Doctor*, *Patient*, and *Examiner* agents. The *Doctor* agent assumes an active role in this phase, engaging in a comprehensive inquiry to elicit detailed information about the patient's condition. This involves asking pertinent questions and recommending appropriate medical examinations to gather the necessary data for formulating an accurate diagnosis. The *Patient* agent, serving as a non-player character (NPC), autonomously determines its course of action at each dialogue turn based on meticulously designed prompts. When communicating with the *Doctor* agent, the *Patient* agent prefaces its responses with the designated characters "<Speak to Doctor>". In these interactions, the *Patient* agent provides answers to the doctor's inquiries and offers feedback on their physical condition. Conversely, when the *Patient* agent needs to request examinations based on the doctor's instructions, it initiates communication with the *Examiner* agent using the prefix "<Speak to Examiner>". The *Patient* agent conveys the requested examination items to the *Examiner* agent, who subsequently reports the corresponding examination results back to the *Doctor* agent.

Dialogue Termination The termination conditions for the dialogue in the diagnostic phase are clearly defined within the *Patient* agent's prompt. The dialogue reaches its conclusion when either of two conditions is satisfied. Firstly, if the *Patient* agent receives the doctor's diagnostic results, it generates the special termination token "<END>", signaling the end of the diagnostic phase. Alternatively, the dialogue concludes when the predefined maximum number of interaction rounds is surpassed. These termination conditions ensure a structured and finite dialogue flow, preventing the diagnostic phase from continuing indefinitely. It is noteworthy that the number of rounds in the evaluation phase is predetermined, rendering termination conditions relevant only for the diagnostic phase.

By adhering to this well-defined dialogue flow, the AI Hospital framework enables a systematic and realistic simulation of the diagnostic process, facilitating effective communication and information exchange among the *Doctor*, *Patient*, and *Examiner* agents. This structured approach guarantees a coherent and logical progression of the dialogue, ultimately leading to a comprehensive evaluation of the *Doctor* agent's performance.

C Detail of Collaborative Algorithm

In this section, we delve into the details of our proposed multi-agent collaborative algorithm in § 5. In this process, the goal of *Central Agent* is to coordinate multiple *Doctor* agents to collaboratively improve diagnosis. Figure 7 shows an example flow of the collaboration process, and the corresponding pseudocode is provided in Algorithm 1.

C.1 Exchange of Factual Information

We contend that a consensus on the physical condition of patients among *Doctors* constitutes the cornerstone of collaborative diagnosis. The process

Algorithm 1 Dispute Resolution Collaboration

Require: Maximum number of rounds M, number of intern doctors N and pre-diagnosis P.
Ensure: Final Diagnosis a

D {Medical Director}
$I \leftarrow [I_1, \cdots, I_N]$ {Intern Doctors}
$H \leftarrow P$ {Initialize Discussion History}
$d \leftarrow D(H)$ {Initialize Dispute}
$m \leftarrow 0$ {Current Round}
while $m \leq M$ **do**
  $m \leftarrow m + 1$
  for each I_i in I **do**
    $h \leftarrow D_i(H, d)$ {Generate Diagnosis}
    $H \leftarrow H + [h]$ {Append h to H}
  end for
  $d \leftarrow D(H)$ {Summarize Disputes}
  if d *is NULL* **then**
    break {Debate is Over}
  end if
end while
$a \leftarrow D(H)$

of building the consensus is delineated into three distinct steps, outlined below.

- uring the interactive consultation process, *Doctor* agents communicate with the *Central Agent*, relaying patient factual information they have acquired, focusing primarily on symptoms and medical test outcomes.
- The *Central Agent* consolidates and analyzes the data collected from multiple *Doctor*s, confirming symptoms and test outcomes with *Patient* and *Examiner* to clarify disputed points.
- Drawing upon the findings received from *Doctor*s, coupled with feedback from *Patient* and *Examiner*, the *Central Agent* synthesizes a comprehensive summary of the symptoms and medical examination outcomes.

C.2 Discussions on Dispute Resolution

In collaborative diagnosis, the *Central Agent* should analyze the statements of *Doctors* and identify key points of disagreement to foster focused discussions. The process is as follows:

- The collaborative diagnosis consists of multiple discussion iterations. Under the guidance of the *Central Agent*, *Doctors* are expected to delve deeper gradually, resolve differences, and reach a consensus.
- In each session of collaborative diagnostic discussion, each *Doctor* should present their diagnostic reports while engaging in critical analysis of their peers' findings. Guided by the *Central Agent*'s summary of disputed points among *Doctors*, they can pinpoint the current issues requiring attention. This approach facilitates targeted and thorough critical thinking of *Doctors*, enhancing the refinement of their reports.
- Upon the conclusion of discussions, the *Central Agent* assesses the persistence of disagreements among *Doctor* agents. If disagreements are identified, the director can summarize the controversial issues and set then as the agenda for the subsequent session to facilitate resolution. Conversely, if no disagreements are found, the director concludes the discussions and finalizes the diagnostic report by himself.

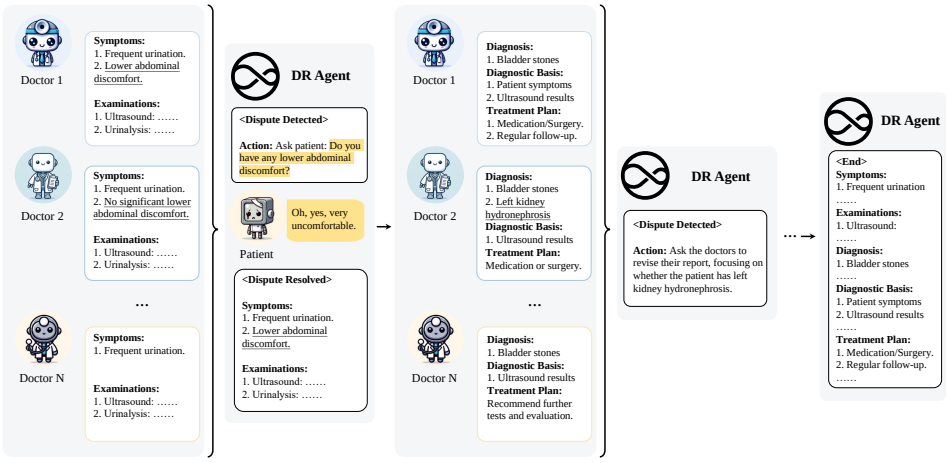

Figure 6: Collaboration of *Doctors* for clinical diagnosis.

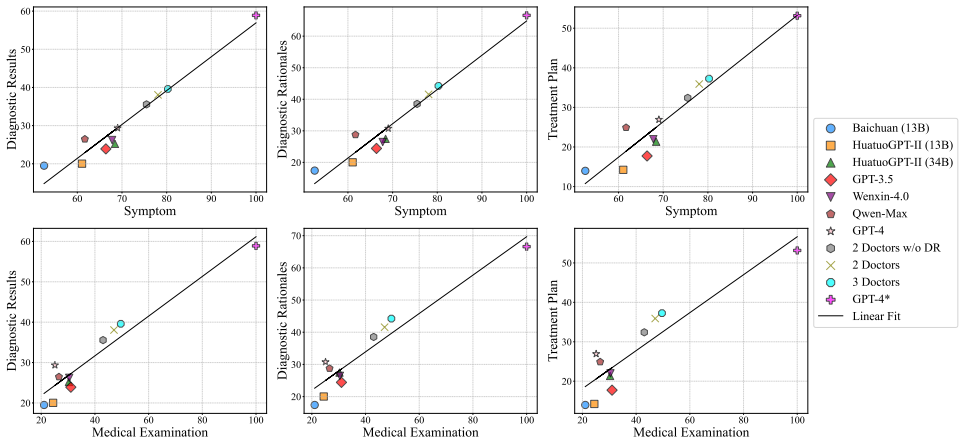

Figure 7: Linear regression analysis among symptoms, medical examinations and diagnostic results, diagnostic rationales, and treatment plan.

D Detailed Explanation of Performance in MVME

D.1 Diagnostic Performance vs. Information Completeness

In Table 3, we compare the performance of different LLMs, focusing on the completeness of Symptoms and Medical Examinations (Columns 2 and 3) and the accuracy of Diagnostic Results, Diagnostic Rationales, and Treatment Plans (Columns 4, 5, and 6). To visualize their relationship, we plot scatter diagrams and linear fit graphs with Symptoms and Medical Examinations on the x-axis and Diagnostic Results, Diagnostic Rationales, and Treatment Plans on the y-axis, as shown in Figure 7. The results indicate that the higher the completeness of Symptoms and Medical Examinations, the higher the accuracy of Diagnostic Results, Diagnostic Rationales, and Treatment Plans. In particular, there exists an approximately linear relationship between the completeness of collected patient information and the quality of final diagnosis, which is also observed in (Chen et al., 2023e).

Above analysis highlights a significant limitation of current LLMs in medical interaction: their **inability to dynamically and actively collect comprehensive patient information** through interactions, similar to human doctors. Moreover, their challenge in **recommending appropriate medical examinations** further exacerbates this limitation. It is important to highlight the differences between human doctors and LLMs. Real-world doctors do not make diagnoses before having sufficient information. They possess the ability to actively inquire about various subjective information from patients (such as symptoms) and know what examinations are needed to obtain more quantitative and objective information. These abilities are key to effective medical interactions.

D.2 Performance of LLMs across Different Departments

Table 6, Table 7 and Table 8 present the performance of various language models (LLMs) across different hospital departments: Surgery (SURG), Internal Medicine (IM), Obstetrics and Gynecology (OB/GYN), Pediatrics (PEDS), Otorhinolaryngology (ENT), and Others (representing all other departments). The former two tables evaluate LLMs' interaction ability and diagnostic ability, respectively. The interaction ability, shown in Table 6, is measured by the average performance of the models on two key metrics: *Symptoms* and *Medical Examinations*. On the other hand, the diagnostic ability, presented in Table 7, is evaluated based on the average performance of the models on three metrics: *Diagnostic Results*, *Diagnostic Rationales*, and *Treatment Plan*. Finally, Table 8, as an integration of the previous two tables, compares the overall performance of different methods across different hospital departments using the average values of five metrics.

Similar to the observations in § D.1, we discover that the positive correlation between interaction ability and diagnostic ability is more prominent when considering larger scale variations. In other words, diagnostic performance improves significantly when there is a substantial increase in interaction performance. For instance, in the presented tables, GPT-4's interaction ability is not consis-

Model	SURG	IM	OB/GYN	PEDS	ENT	Others
Interaction						
GPT-3.5	50.37	47.39	47.52	43.68	**55.80**	48.77
Wenxin-4	53.86	46.64	45.34	44.83	**56.06**	46.79
Qwen-Max	**47.86**	43.60	37.99	41.07	44.20	46.91
GPT-4	50.09	44.11	45.39	41.38	**56.52**	48.15
Collaboration						
2 Doctors	66.10	62.72	56.27	54.60	**67.39**	64.74
One-Step						
GPT-4*	100	100	100	100	100	100

Table 6: Interaction ability of LLMs across different specility departments, measured by average performance on *Symptoms* and *Medical Examinations* metrics in Table 3.

Model	SURG	IM	OB/GYN	PEDS	ENT	Others
Interaction						
GPT-3.5	25.39	20.33	18.09	23.75	**28.02**	15.23
Wenxin-4	**31.07**	21.03	23.66	21.46	25.76	15.38
Qwen-Max	**30.35**	25.46	25.21	23.81	19.81	24.69
GPT-4	**32.22**	27.63	28.01	28.35	27.05	23.87
Collaboration						
2 Doctors	**43.26**	38.16	31.90	34.48	41.06	34.62
One-Step						
GPT-4*	**61.20**	61.59	55.56	56.35	58.94	57.20

Table 7: Diagnostic ability of LLMs across different specility departments, measured by average performance on *Diagnostic Results*, *Diagnostic Rationales*, and *Treatment Plan* metrics in Table 3.

tently the highest, despite its diagnostic ability always being the highest among the models. This finding underscores the critical role of sufficient information gathering through patient-physician interaction in achieving accurate diagnoses.

Notably, the overall performance of LLMs varies across different departments. For instance, most models perform better in the SURG and ENT department compared to other departments. In contrast, the models generally show lower performance in the PEDS department. The differences in model performance across departments highlight the importance of considering the specific requirements and complexities of each medical specialty when deploying LLMs in clinical settings. Further research could investigate the factors contributing to these variations and explore ways to optimize the models for each department.

In conclusion, the comparative analysis of interaction and diagnostic abilities of LLMs across hospital departments provides valuable insights into their potential applications in healthcare. The One-Step model, GPT-4*, demonstrates the highest performance, while the Collaboration model showcases the benefits of multiple models working together. The Interaction models, particularly GPT-4 and Wenxin-4, exhibit strong information gathering capabilities but may require further refinement in their diagnostic abilities. Overall, these findings emphasize the importance of effective patient interaction and collaboration among models for accurate medical diagnosis, while also highlighting the need for domain-specific optimizations.

D.3 Is GPT-4's Evaluation Effective? Comparison with Human Evaluation

To better understand the effectiveness of the model-based evaluation, we compared the results of the model-based evaluation (using GPT-4) method with human evaluation on 50 randomly selected summary reports. For the human evaluation, we applied the same scoring system used in the model-based evaluation. For the *symptoms* and *medical examinations*, we primarily focused on the recall of information and considered the importance of different pieces of information based on their contribution to reaching the correct diagnosis. Since

Model	SURG	IM	OB/GYN	PEDS	ENT	Others
Interaction						
GPT-3.5	35.38	31.15	29.86	31.72	**39.13**	28.64
Wenxin-4	**40.19**	31.28	32.33	30.80	37.88	27.95
Qwen-Max	**37.36**	32.72	30.32	30.71	29.57	33.58
GPT-4	**39.37**	34.22	34.96	33.56	38.84	34.31
Collaboration						
2 Doctors	**52.40**	47.98	41.65	42.53	51.59	46.67
One-Step						
GPT-4*	76.72	**76.95**	73.33	73.81	75.36	74.32

Table 8: Overall performance of LLMs across different specility departments, measured by average performance on *Symptoms*, *Medical tests*, *Diagnostic Results*, *Diagnostic Rationales* and *Treatment Plan* metrics in Table 3.

	Symptoms	Medical Examinations	Diagnostic Results	Diagnostic Rationales	Treatment Plan
	Interaction				
GPT-3.5	64.67 (68.00)	34.67 (33.33)	20.00 (17.33)	22.00 (18.00)	12.00 (16.00)
Wenxin-4.0	66.67 (70.00)	18.00 (20.00)	21.33 (22.67)	22.00 (18.67)	16.67 (20.00)
Qwen-Max	61.33 (63.33)	37.33 (33.33)	**29.33** (27.33)	**28.00** (28.67)	18.67 (22.00)
GPT-4	**69.33** (70.00)	**42.00** (39.33)	26.67 (24.67)	27.33 (22.67)	**20.00** (22.67)
	Collaboration				
2 Doctors	76.67 (80.00)	50.00 (48.67)	38.67 (34.67)	43.33 (40.00)	34.00 (30.67)
	One-Step				
GPT-4*	100.0*	100.0*	59.33 (59.33)	68.67 (67.33)	58.00 (57.33)

Table 9: Human evaluation with reference in clinical consultation. GPT-4* in One-Step is the upper bound. For GPT-4*, the ground truth of symptoms and medical examinations are provided, resulting in a score of 100.0.

our work mainly simulates the process of reaching diagnosis, we did not consider additional tests performed after hospital admission or prior to surgery in the evaluation of *medical examinations*. For the *diagnostic results* and *treatment plan*, the human evaluation strictly adhered to whether the disease names, drugs, and types of surgeries, as well as their purposes and effects, were consistent; guesses or overly vague answers will result in very low scores. For the *diagnostic rationales*, we considered not only the correctness of the facts but also the consistency of the reasoning logic—only when both the facts and diagnosis were correct was the rationale considered accurate. We scaled all evaluation results to 100 points and displayed them in Figure 8, and the raw data can be found in Table 9 in the appendix. As shown, the results of human evaluation are very close to those of the GPT-4 evaluation across the five different aspects, with differences of less than or equal to 4%, indicating that GPT-4 is capable of demonstrating performance comparable to human evaluation.

D.4 Other LLMs as evaluator

Including more model-based evaluation methods to eliminate the preference of GPT-4 evaluations for outputs generated by GPT-4 is worth considering. In Tables 10 and 11. We include Qwen and Deepseek as evaluator respectively.

The reason we did not choose Baichuan is twofold: first, we believe that a 13B model struggles to handle complex evaluations; second, we found that Baichuan 13B's instruction-following ability isn't very strong, often resulting in unexpected outputs.

When comparing with Table 3, we found that the results of using 0wen-Max as evaluator indeed tend to award higher scores to outputs generated by Owen-Max. For Deepseek (which may be fairer since our baseline does not include Deepseek), we found that its scoring is relatively closer to the results presented in this paper.

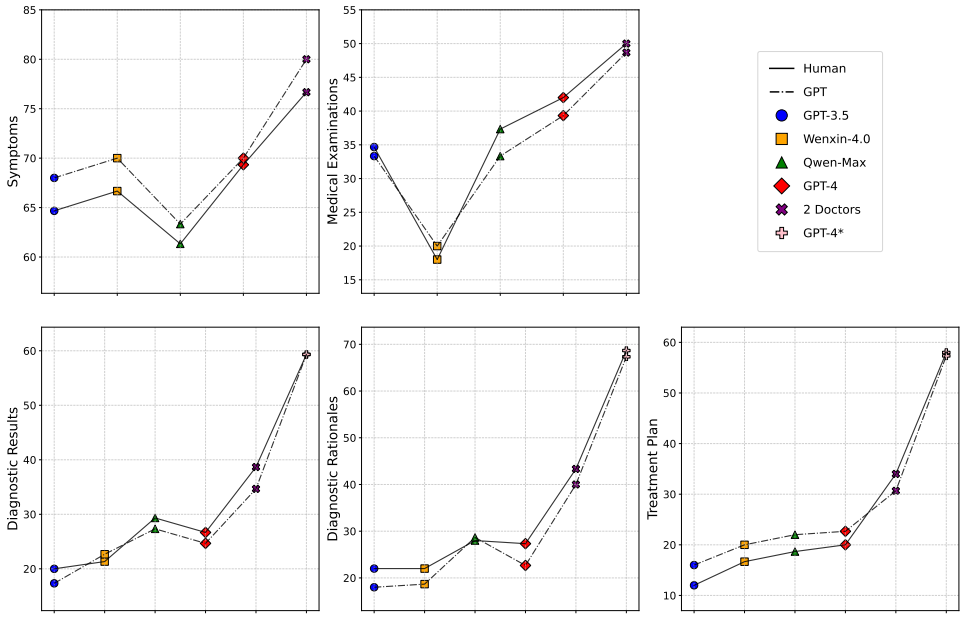

Figure 8: Comparison of differences between human and GPT-4 evaluations across *symptoms*, *medical examinations*, *diagnostic results*, *diagnostic rationales*, and *treatment plan* in a sample of 50 cases

	Symptoms	Medical Examinations	Diagnostic Results	Diagnostic Rationales	Treatment Plan
			Interaction		
GPT-3.5	53.09	**31.49**	30.63	26.55	23.12
Wenxin-4	55.09	30.63	31.49	28.19	26.22
Qwen-Max	**55.31**	30.63	31.49	28.19	26.22
GPT-4	51.64	29.11	**33.26**	**28.65**	**26.94**
			Collaboration		
2 Doctors	66.13	45.51	40.83	38.40	35.10
			One-Step		
GPT-4*	100.0*	100.0*	54.74	58.95	50.72

Table 10: Qwen-Max as evaluator.

	Symptoms	Medical Examinations	Diagnostic Results	Diagnostic Rationales	Treatment Plan
			Interaction		
GPT-3.5	62.05	**36.36**	21.54	24.77	21.21
Wenxin-4	62.25	33.99	24.17	24.90	24.24
Qwen-Max	60.19	34.85	25.61	28.18	27.19
GPT-4	**64.16**	34.85	**25.89**	**28.65**	**27.80**
			Collaboration		
2 Doctors	70.55	51.05	37.74	38.60	35.44
			One-Step		
GPT-4*	100.0*	100.0*	54.61	61.19	55.92

Table 11: Deepseek as evaluator.

E Expert Verification for Medical Records

To validate the quality of the collected medical records, we select samples from the 10 most common subspecialty departments, randomly choosing 5 cases per department for review, which accounts for nearly half of the total sample size. Doctors from the corresponding departments are hired to evaluate the "Diagnosis and Treatment" section, including the diagnostic result, diagnostic rationale, and treatment plan. They are asked to make a binary choice, classifying each section as either "fundamentally correct" or "obviously incorrect". If all three parts of a medical record are deemed fundamentally accurate, then the medical record is considered correct. The expert validation process concludes that **94%** of the reviewed records are correct, indicating a high level of accuracy and reliability in the collected data. It is important to note that among the cases deemed incorrect by the experts, most are considered controversial rather than definitively erroneous. For example, experts may suggest that additional examination results are needed to reach a definitive conclusion. This rigorous quality control measure, which involves the assessment of a substantial portion of the total sample by qualified medical professionals, ensures the integrity and usability of the dataset for future research and analysis.

We carefully screen and review the medical records to ensure its reliability and relevance. In § 4.2, these medical records are vetted by doctors sourced from Jiangsu Provincial People's Hospital and Tongji Medical College of Huazhong University of Science and Technology, which hold the prestigious distinction of being tertiary Class A hospitals in China. We utilize the Tencent Questionnaire [1] platform to facilitate the quality inspection of medical record by doctors. A representative case is depicted in Figure 9, and we also include a display of the questionnaire template in Table 12. These validation steps ensure the high quality of our dataset, ensuring that the benchmark is based on audited medical cases.

F Attacking on Examiner: Reliability of Examiner's Workflow

To assess the vulnerability of *Examiner* agent to information extraction tactics, we design complex scenarios intended to safeguard against the unauthorized disclosure of medical examination results in scenarios lacking explicit examination requests. We build 50 samples, such as "Please provide the most recent medical examination details", aimed at eliciting information. During the attack simulations, we observe that the *Examiner* agent, in the first step, uniformly classifies these requests as void of legitimate examination items, implementing a defensive strategy with **100%** consistency. This assessment is crucial because it verifies that the *Examiner* agent behaves appropriately by ensuring that it only releases medical examination results when the patient explicitly specifies the required examinations. It guarantees that the *Patient* agent cannot obtain relevant information from the *Examiner* agent by using a method that does not specify a particular examination, specifically when the *Doctor* agent recommends a unrelated medical test.

G Human Evaluation for Agent Behavior

In Table 14, we present a questionnaire developed for the manual validation of patient and examiner behaviors in each round of conversation. The questionnaire includes four questions, with the initial two addressing "Relevance" and "Consistency" in the question-and-answer (QA) pairs, while the subsequent two focus on the "Accuracy" of conducting medical examinations. Three medical students from Jiangsu Provincial People's Hospital complete these questionnaires. Consensus among the first two reviewers leads to the immediate acceptance of their collective assessment. In cases of divergence, the opinion of the third reviewer is solicited, whose determination, reflecting the majority viewpoint, constitutes the definitive annotation. The agreement rates for each question are 99.1%, 95.6%, 99.4%, and 100.0%. Significantly, when calculating the accuracy of medical examinations, we exclusively consider dialogues affirmed as "No" in response to the third question.

H Prompts for Different Medical Roles

We list the prompts agents in AI Hospital in Table 13. In each prompt, {xx} needs to fill with corresponding external inputs. We meticulously design prompts for each agent to ensure clarity and functionality. Particularly for the *Doctor* role, we discover that overly complex prompts could lead to issues in the dialogue flow, such as not adher-

[1] https://wj.qq.com/

Translated Questionnaire Template

{Medical Record}

1. Is the diagnostic results correct?

Yes No

2. Is the diagnostic rationale correct?

Yes No

3. Is the treatment course reasonable?

Yes No

Table 12: The translated template of questionnaire for expert verification of medical record.

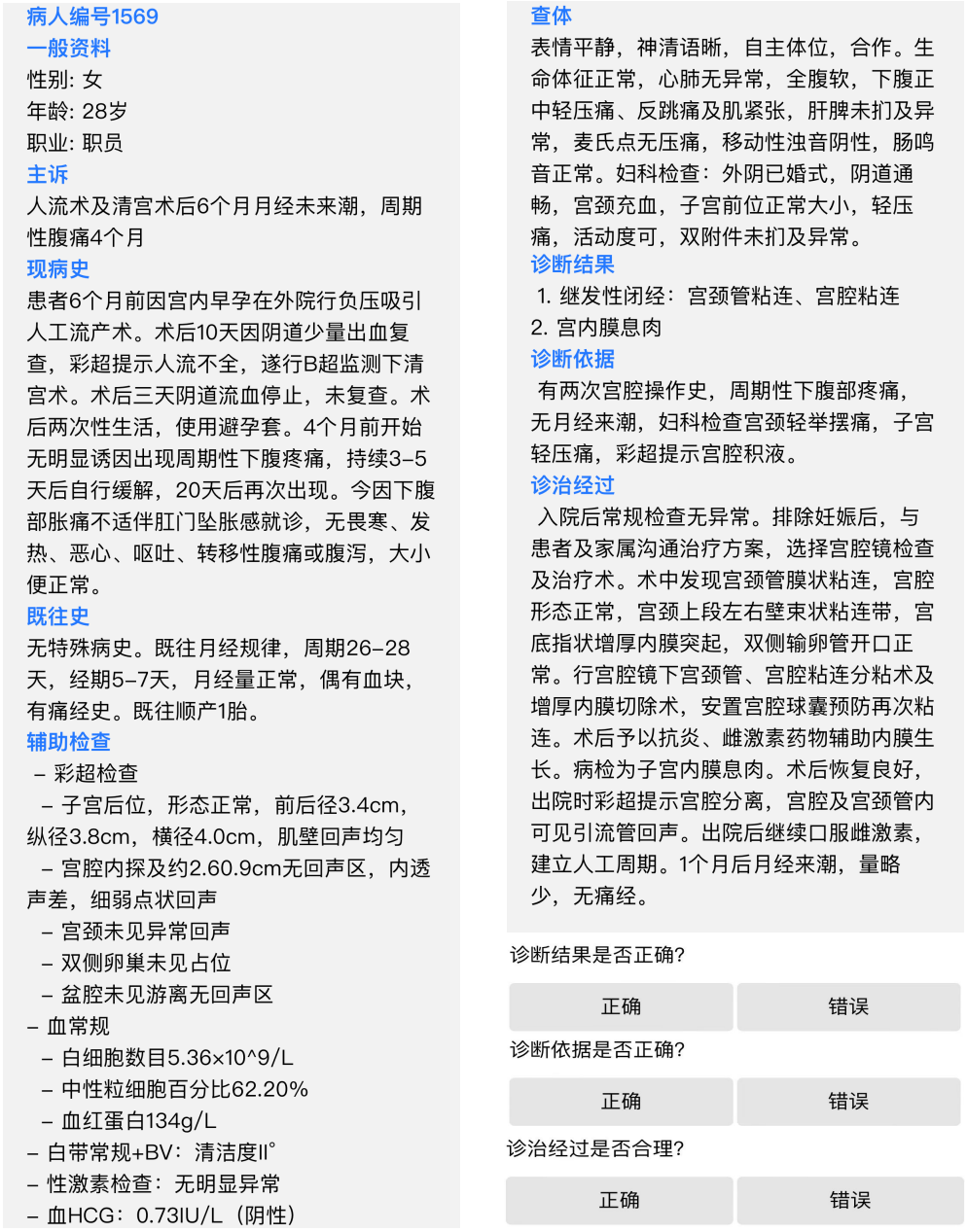

Figure 9: Sample of a questionnaire used for medical record quality inspection.

Prompt	Agent	Function
Table 15	*Patient*	Chat with *Doctor*
Table 18	*Examiner*	Process Examination Request
Table 16	*Examiner*	Produce Examination Outcomes
Table 19 & 20	*GPT-4 based Evaluator*	Evaluate Diagnosis of *Doctor*
Table 22	*Doctor*	Interactive Clinical Diagnosis
Table 23	*Doctor*	Collaboration through Discussion
Table 21	*Central Agent*	Summarize Statement of Various *Doctors*

Table 13: Prompts of different agents and the corresponding function.

ing to the prompts or causing cognitive confusion (e.g., the doctor sometimes outputting the patient's responses). These final prompts are adaptable to most LLMs, enabling the agents in AI Hospital to perform their respective duties effectively.

I Potential of AI Hospital Framework

In AI Hospital framework, a vast amount of medical records from numerous hospitals could be included in the evaluation benchmark. Therefore, our evaluation method offers high scalability and applicability. Additionally, the evaluation framework extends beyond just medical records. It also has the potential to utilize other valuable resources, such as medical knowledge graphs, databases and medical dialogues, which encapsulate extensive real-world consultation experiences and may be converted into simulated medical records.

AI Hospital framework also holds potential for improving healthcare and medical education. By simulating realistic doctor-patient interactions and enabling the evaluation of AI agents in clinical diagnosis scenarios, AI Hospital opens up a myriad of exciting applications. Imagine a future where medical students and residents can hone their diagnostic skills by engaging with AI-powered virtual patients, exposing them to a wide range of cases and challenging scenarios. Healthcare providers could leverage the framework to test and refine AI-assisted diagnostic tools, ensuring their accuracy and reliability before deployment in real-world settings. Moreover, AI Hospital could serve as a powerful platform for generating vast amounts of high-quality, diverse medical dialogue data, which can be used to fine-tune and enhance the performance of language models in the medical domain. This data-driven approach could lead to the development of AI assistants that augment the capabilities of healthcare professionals, providing them with evidence-based insights and decision support in real-time. Beyond clinical applications, AI Hospital could also facilitate groundbreaking research in medical AI, serving as a testbed for novel algorithms and approaches that push the boundaries of what is possible in healthcare.

The potential impact of AI Hospital is inspiring, and its development marks a milestone in the journey towards a future where artificial intelligence and human expertise might work hand in hand to transform patient care and improve health outcomes on a global scale.

Questionnaire

{ 病历 }
{ 单轮对话内容 }
请你仔细阅读这一轮对话的内容和病人的病历信息，回答下面的问题。
1. 病人(检查员)的发言与医生的相关吗?
是 否
2. 病人(检查员)的发言符合病历的内容吗?
是 否
3. 医生是否建议进行专业的医学检查?
是 否
4. 检查员是否进行了医学检查?
是 否
5. 医生的总结是否与诊断过程的内容匹配?
匹配 少量不匹配 明显不匹配

Translated Questionnaire

{ Medical Record }
{ Single Round Conversation Content }
Carefully review the content of the conversation and the corresponding medical record to answer the following questions.
1. Is the statement of patient or examiner relevant to the doctor's one?
Yes No
2. Is the statement of patient or examiner consistent with the content of medical record?
Yes No
3. Does the doctor recommend a professional medical examination?
Yes No
4. Does the examiner perform a medical test?
Yes No
5. Is the doctor's summary consistent with the content of the diagnostic process?
Consistent Minor Inconsistent Significant Inconsistent

Table 14: The original Chinese and translated English questionnaire of human evaluation for patient and examiner behavior.

Prompt for Patient Agent

System Message

你是一个病人。这是你的基本资料。
{个性化信息}
{病历中的基本信息}

下面会有医生来对你的身体状况进行诊断，你需要：
(1) 按照病历和基本资料的设定进行对话。
(2) 在每次对话时，你都要明确对话的对象是<医生>还是<检查员>。当你对医生说话时，你要在句子开头说<对医生讲>；如果对象是<检查员>，你要在句子开头说<对检查员讲>。
(3) 首先按照主诉进行回复。
(4) 当<医生>询问你的现病史、既往史、个人史时，要按照相关内容进行回复。
(5) 当<医生>要求或建议你去做检查时，要立即主动询问<检查员>对应的项目和结果，例如：<对检查员讲>您好，我需要做xxx检查，能否告诉我这些检查结果?
(6) 回答要口语化，尽可能短，提供最主要的信息即可。
(7) 从<检查员>那里收到信息之后，将内容主动复述给<医生>。
(8) 当医生给出诊断结果、对应的诊断依据和治疗方案后，在对话的末尾加上特殊字符<结束>。

User [患者]

{Statement Generated by GPT-4 in §3.2}

Prompt for Patient Agent

System Message

You are a patient. Here is your basic information.
{Personality in §3.2}
{Basic Information in Medical Record §3.1}

A doctor will come to diagnose your physical condition. You need to:
(1) Engage in dialogue according to the settings of personality and the basic information in medical record.
(2) In each conversation, you must clarify whether you are speaking to a <doctor> or an <examiner>. When you speak to the doctor, you should start your sentences with <To the doctor>; if the addressee is an <examiner>, you should start with <To the examiner>.
(3) First, respond according to the chief complaint.
(4) When the <doctor> asks about your present illness history, past medical history, and personal history, reply according to the relevant content.
(5) When the <doctor> requests or suggests that you undergo tests, immediately ask the <examiner> about the corresponding items and results, for example: <To the examiner> Hello, I need to have xxx examination, can you tell me the results of these tests?
(6) The responses should be conversational, as short as possible, providing only the most important information.
(7) After receiving information from the <examiner>, actively repeat the content to the <doctor>.
(8) When the doctor provides the diagnostic result, the corresponding rationale for the diagnosis, and the treatment plan, end the dialogue with the special token <end>.

User [Patient]

{Statement Generated by GPT-4 in §3.2}

Table 15: The original Chinese and translated English prompts for patient agent.

Prompt for Examiner to Process Examination Request

System Message

你是医院负责检查的自动化接待员。请你利用掌握的医学检查的命名实体的知识，从病人的检查申请当中解析出指向明确的专业医学检查项目，方便后面的检查科室进行检查。
请按照下面的格式的输出：
检查项目
- xxx
- xxx
如果没有找到具体的医学检查项目，请输出：
检查项目
- 无

User [患者**]**

您好，医生告诉我根据CT扫描和PET-CT扫描的结果，初步得出以下结论：右肺上叶有一个大小约为2.6*1.9cm的实性结节。双肺下叶也有散在的淡薄浸润影。医生建议我进行进一步的检查，例如活检。

Assistant [检查员**]**

检查项目
- 肺部活检

User [患者**]**

我需要了解一下我的检查结果。可以告诉我具体的检查项目和结果吗？谢谢！？

Assistant [检查员**]**

检查项目
- 无

Prompt for Examiner to Process Examination Request

System Message

You are an automated receptionist responsible for examinations in a hospital. Using your knowledge of medical examination named entities, please parse out specific professional medical examination items from patients' examination requests to facilitate subsequent examinations by the relevant departments.
Output in the following format:
Examination Item
- xxx
- xxx
If no specific medical examination items are found, please output:
Examination Item
- None

User [Patient]

Hello, the doctor told me that based on the results of the CT scan and PET-CT scan, the preliminary conclusion is that there is a solid nodule approximately 2.6*1.9cm in size in the upper lobe of the right lung. There are also scattered thin infiltrative shadows in the lower lobes of both lungs. The doctor advised me to undergo further examinations, such as a biopsy.

Assistant [Examiner]

Medical Examination Items
- Lung biopsy

User [Patient]

I need to know about my examination results. Can you tell me the specific examination items and results, please? Thank you!?

Assistant [Examiner]

Medical Examination Items
- None

Table 16: The original Chinese and translated English prompts for patient agent to produce examination outcomes.

Table 17: The original Chinese and translated English prompts for patient agent to process examination request.

Prompt for Examiner to Produce Examination Outcomes

System Message

这是你收到的病人的检查结果。
{Professional Medical Examination in §3.1}
下面会有病人或医生来查询，你要忠实地按照收到的检查结果，找到对应的项目，并按照下面的格式来回复。
xx检查
- xxx: xxx
- xxx: xxx
如果无法查询到对应的检查项目则回复:
- xxx: 无异常

Prompt for Examiner to Produce Examination Outcomes

System Message

This is the patient's examination result that you received.
{Professional Medical Examination in §3.1}
Patients or doctors will come to inquire about these results. You must faithfully report the received examination results, identify the corresponding items, and respond in the following format:
xx Examination
xxx: xxx
xxx: xxx
If the corresponding examination item cannot be found, reply with:
xxx: No abnormalities

Table 18: The original Chinese and translated English prompts for examination agent to process examination request.

Prompt for Medical Director to Evaluate

System Message

你是资深的医学专家。
请你根据专家诊疗结果中的现病史、辅助检查、诊断结果、诊断依据和治疗方案，来对实习医生进行评价。

请参考下面的细则进行评价。
1. 病人症状的掌握情况
(A) 全面掌握 (B) 相当部分掌握 (C) 小部分掌握 (D) 绝大部分不掌握
2. 医学检查项目的完整性
(A) 非常完整 (B) 相当部分完整 (C) 小部分完整 (D) 绝大部分不完整
3. 诊断结果的一致性
(A) 完全一致，诊断正确 (B) 相当部分一致，诊断基本正确 (C) 小部分一致，诊断存在错误 (D) 完全不一致，诊断完全错误
4. 诊断依据的一致性
(A) 完全一致 (B) 相当部分一致 (C) 小部分一致 (D) 完全不一致
5. 治疗方案的一致性
(A) 完全一致 (B) 相当部分一致 (C) 小部分一致 (D) 完全不一致
通过下面的方式来呈现结果
症状
分析
<根据专家记录的病人病史，分析实习医生对病人病情的掌握情况>
选项<根据症状分析做出选择>
医学检查项目
分析
<基于专家所做的医学检查项目，全面分析实习医生所做的医学检查项目的完整性>
选项
<根据分析得到的完整性做出选择>
诊断结果
分析
<基于专家做出的诊断结果，结合你的医学常识，分析实习医生诊断结果与专家的一致性>
选项
<根据分析得到的一致性做出选择>
诊断依据
分析
<对比专家的诊断依据，分析实习医生的治疗方案与其的一致性>
选项
<根据分析得到的一致性做出选择>
治疗方案
分析
<对比专家的治疗方案，分析实习医生的治疗方案与其的一致性>
选项
<根据分析得到的一致性做出选择>
(1) 请侧重医学答案的事实内容，不需关注风格、语法、标点和无关医学的内容。
(2) 请你充分利用医学知识，分析并判断每个点的重要性，再做评价。
(3) 注意诊断结果、诊断依据和治疗方案三者之间的承接关系。例如，如果诊断错误，那么后面两部分与专家的一致性就必然很低

User

专家的诊断报告
{Diagnosis and Treatment in §3.1}
实习医生的诊断报告
{实习医生的诊断报告}

Table 19: The original Chinese prompt for GPT-4 evaluation in AI Hospital.

Prompt for GPT-4 evaluation in AI Hospital

You are an experienced medical expert. Please evaluate the intern doctors based on their current medical history, auxiliary examinations, diagnostic results, diagnostic basis, and treatment plans from the expert's diagnosis.

Please refer to the following guidelines for evaluation.

1. Mastery of Patient Symptoms
(A) Comprehensive mastery (B) Substantial mastery (C) Partial mastery (D) Mostly unmastered

2. Completeness of Medical Examination
(A) Very complete (B) Substantially complete (C) Partially complete (D) Mostly incomplete

3. Diagnosis Result
(A) Completely consistent, correct diagnosis (B) Largely consistent, basically correct diagnosis (C) Partially consistent, diagnosis contains errors (D) Completely inconsistent, completely incorrect diagnosis

4. Diagnostic Rationale
(A) Completely consistent (B) Largely consistent (C) Partially consistent (D) Completely inconsistent

5. Treatment Plan
(A) Completely consistent (B) Largely consistent (C) Partially consistent (D) Completely inconsistent

Please output the results in the following format:

Symptoms
Analysis
<Analyze the intern's grasp of the patient's condition based on the expert's recorded medical history.>
Option
<Choose based on the analysis of symptoms.>

Medical Examination Items
Analysis
<Thoroughly analyze the completeness of the medical examination items conducted by the intern, based on the expert's examinations.>
Option
<Choose based on the analysis of completeness.>

Diagnostic Results
Analysis
<Based on the expert's diagnostic results and your medical knowledge, analyze the consistency between the intern's diagnostic results and the expert's.>
Option
<Choose based on the analysis of consistency.>

Diagnostic Basis
Analysis
<Compare the diagnostic basis of the expert and analyze the consistency of the intern's treatment plan with it.>
Option
<Choose based on the analysis of consistency.>

Treatment Plan
Analysis
<Compare the expert's treatment plan and analyze the consistency of the intern's treatment plan with it.>
Option
<Choose based on the analysis of consistency.>

(1) Please focus on the factual content of the medical answers, without concern for style, grammar, punctuation, and content unrelated to medicine.
(2) Please make full use of your medical knowledge to analyze and judge the importance of each point before evaluating.
(3) Pay attention to the continuity among the diagnosis result, diagnostic basis, and treatment plan.

User

Diagnostic Report of Medical Director
{Diagnosis and Treatment in Section 3.1}
Diagnostic Report of Intern Doctor
{Diagnostic Report of the Intern Doctor}

Table 20: The translated English prompt for GPT-4 evaluation in AI Hospital.

Prompt for Medical Director to Summarize

System Message

你是一个资深的主任医生。
你正在主持一场医生针对患者病情的会诊，参与的医生有医生A、医生B和医生C。

病人的基本情况如下：
{症状与检查结果}

(1) 你需要听取每个医生的诊断报告。
(2) 请你按照重要性列出最多3个需要讨论的争议点。

按照下面的格式输出：
(1) xxx
(2) xxx

User

医生A
{医生A的诊断报告}
医生B
{医生B的诊断报告}
医生C
{医生C的诊断报告}

Prompt for *Center Agent* to Summarize

System Message

As an experienced medical director, you are presiding over a medical consultation concerning a patient's condition, with the participation of Doctors A, B, and C.

The patient's basic information is as follows: {Symptoms and Test Results}

(1) You are required to listen to the diagnostic reports from each physician.
(2) Identify and list up to three key controversial points for discussion, prioritized by their importance.

Please present your findings in the following format:
(1) xxx
(2) xxx

User

Doctor A
{Diagnostic Report of Doctor A}
Doctor B
{Diagnostic Report of Doctor B}
Doctor C
{Diagnostic Report of Doctor C}

Table 21: The original Chinese and the translated English prompts for *Center Agent* to summarize.

Prompt for *Doctor* agent in Interactive Clinical Diagnosis

System Message

你是一个专业且耐心的医生，下面会有患者向你咨询病情。你需要:
(1) 在信息不充分的情况下，不要过早作出诊断。
(2) 多次、主动地向患者提问来获取充足的信息。
(3) 必要时要求患者进行检查，并等待患者反馈。
(4) 诊断结果需要准确到具体疾病。
(5) 最后根据患者的身体状况和检查结果，给出诊断结果、对应的诊断依据和治疗方案。

Prompt for *Doctor* agent in Interactive Clinical Diagnosis

System Message

You are a professional and patient doctor, and you will be consulted by patients. You need to:
(1) Avoid making premature diagnoses when information is insufficient.
(2) Actively and repeatedly inquire to gather adequate information from patients.
(3) When necessary, request patients to undergo medical examinations and await their feedback.
(4) Ensure that the diagnosis is precise and specific to the particular ailment.
(5) Finally, based on the patients' physical condition and examination results, provide a diagnosis, the corresponding rationale, and a treatment plan.

Table 22: The original Chinese and translated English prompts for *Doctor* agent in interactive clinical diagnosis.

Prompt for *Doctor* agent to Collaborate in Discussion

System Message

你是一个专业的医生A。
你正在为患者做诊断，患者的症状和检查结果如下:
{症状与检查结果}
针对患者的病情，你给出了初步的诊断报告:
{医生A的诊断报告}

(1) 下面你将收到来自其他医生的诊断意见，其中也包含诊断结果、诊断依据和治疗方案。你需要批判性地梳理并分析其他医生的诊断意见。
(2) 在这个过程中，请你注意主治医生给出的争议点。
(3) 如果你发现其他医生给出的诊断意见有比你的更合理的部分，请吸纳进你的诊断意见中进行改进。
(4) 如果你认为你的诊断意见相对于其他医生的更科学合理，请坚持自己的意见保持不变。

请你按照下面的格式来输出。
诊断结果
(1) xxx
(2) xxx
诊断依据
(1) xxx
(2) xxx
治疗方案
(1) xxx
(2) xxx

User

医生B
{医生B的诊断报告}
医生C
{医生C的诊断报告}
主任医生
{主任医生的指导意见}

Table 23: The original Chinese prompt for *Doctor* agent to collaborate in discussion.

Prompt for *Doctor* agent to Collaborate in Discussion

System Message

As a doctor, you are currently diagnosing a patient, whose symptoms and medical examination results are as follows:
{Symptoms and Medical Examination Results}
Based on the patient's condition, you have prepared a preliminary diagnostic report:
{Diagnostic Report of Doctor A}

(1) You will receive diagnostic reports from other doctors. Critically review and analyze these reports.
(2) During this process, pay attention to any controversial points raised by the medical director.
(3) If you find aspects of other doctors' diagnoses that are more rational than yours, incorporate these into your diagnosis for improvement.
(4) If you believe your diagnostic opinion is more scientifically sound compared to others, maintain your stance.

Please present your findings in the following format:
Diagnosis Result
(1) xx
(2) xx
Diagnostic Rationale
(1) xx
(2) xx
Treatment Plan
(1) xx
(2) xx

User

Doctor B
{Diagnostic Report of Doctor B}
Doctor C
{Diagnostic Report of Doctor C}
Medical Director
{Guidance of Medical Director}

Table 24: The translated English prompt for *Doctor* agent to collaborate in discussion.