GPBench: A Comprehensive and Fine-Grained Benchmark for Evaluating Large Language Models as General Practitioners

Zheqing Li[1*] **Yiying Yang**[2*] **Jiping Lang**[1*] **Wenhao Jiang**[2*†‡]
Yuhang zhao[2] **Shuang Li**[1] **Dingqian Wang**[2] **Zhu Lin**[1] **Xuanna Li**[1]
Yuze Tang[1] **Jiexian Qiu**[3] **Xiaolin Lu**[3] **Hongji Yu**[3] **Shuang Chen**[1]
Yuhua Bi[1] **Xiaofei Zeng**[1] **Yixian Chen**[4] **Junrong Chen**[1†] **Lin Yao**[1†‡]
[1] The Sixth Affiliated Hospital of Sun Yat-sen University
[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[3] Xinyi People's Hospital
[4] School of Intelligent Systems Engineering, Sun Yat-sen University
cswhjiang@gmail.com, {chenjr5,yaolin}@mail.sysu.edu.cn

Abstract

General practitioners (GPs) serve as the cornerstone of primary healthcare systems by providing continuous and comprehensive medical services. However, due to community-oriented nature of their practice, uneven training and resource gaps, the clinical proficiency among GPs can vary significantly across regions and healthcare settings. Currently, Large Language Models (LLMs) have demonstrated great potential in clinical and medical applications, making them a promising tool for supporting general practice. However, most existing benchmarks and evaluation frameworks focus on exam-style assessments—typically multiple-choice question—lack comprehensive assessment sets that accurately mirror the real-world scenarios encountered by GPs. To evaluate how effectively LLMs can make decisions in the daily work of GPs, we designed GPBench, which consists of both test questions from clinical practice and a novel evaluation framework. The test set includes multiple-choice questions that assess fundamental knowledge of general practice, as well as realistic, scenario-based problems. All questions are meticulously annotated by experts, incorporating rich fine-grained information related to clinical management. The proposed LLM evaluation framework is based on the competency model for general practice, providing a comprehensive methodology for assessing LLM performance in real-world settings. As the first large-model evaluation set targeting GP decision-making scenarios, GPBench allows us to evaluate current mainstream LLMs. Expert assessment and evaluation reveal that in areas such as disease staging, complication recognition, treatment detail, and medication usage, these models exhibit at least ten major shortcomings. Overall, existing LLMs are not yet suitable for independent use in real-world GP working scenarios without human oversight.

***Keywords*—** Large Language Models, Benchmark, General practitioners, Competency Model

[*] Co-first author.
[†] Corresponding authors.
[‡] Co-lead the project.

Preprint. Under review.

1 Introduction

Large Language Models (LLMs) have emerged as a prominent technology in recent years. By leveraging vast amounts of data and substantial computational power, these models often trained on publicly available online text—have grown exceedingly large. In general, the bigger the model, the more data it uses, and the more compute resources it consumes, the stronger its capabilities become. LLMs have already shown impressive advances in mathematical reasoning, dialogue, coding, and other areas. Because of their extensive knowledge bases and reasoning abilities, they also hold great promise for applications in healthcare.

As model sizes continue to scale up, for instance, MedPaLM [1], MedPaLM 2 [2], and GPT-4 [3], these models claim to deliver performance on certain medical tasks that is approaching or even surpassing that of human experts. GPT-4, for example, outperforms medical students in open-ended clinical reasoning exams, particularly for complex cases and in the "list the problems" task [4]. In [5], researchers observed that ChatGPT performs at a level comparable to a third-year medical student on assessments of primary medical knowledge. Med-PaLM [1] was the first model to achieve a passing score on the MedQA dataset [6], which comprises questions in the style of the U.S. Medical Licensing Examination (USMLE). According to [7], GPT-4—even without specialized prompt engineering—exceeds the USMLE passing score by over 20 points. By improving domain-specific fine-tuning and adopting novel prompting strategies, Med-PaLM2 [2] further boosts performance on medical question-answering tasks, attaining a score of 86.5% on the MedQA [6] dataset.

Existing LLMs can analyze patient symptom descriptions and medical histories, assist physicians in making more accurate diagnoses, and generate a wealth of study materials to help medical students understand and master complex medical knowledge. In addition, LLM-based conversational systems can answer frequent questions from patients, offer health advice, and guide self-management, thereby significantly enhancing patient experience and the overall quality of healthcare services.

Despite the enormous potential of LLMs in medical decision support systems [8, 9, 10, 11], current evaluations of their performance focus primarily on knowledge-based or specialty-specific tasks [12]. In this paper, we evaluate the capabilities of LLMs in a general practice setting. General practitioners, serving as the "first point of contact" in a tiered healthcare system, handle initial consultations, routine check-ups, diagnostics, and basic treatments for community residents or grassroots populations. In a single visit, they may assess multiple common, frequently occurring, and chronic diseases, while also providing prevention, health maintenance, and essential health education. Previous evaluation methods have often been overly simplistic—typically limited to multiple-choice tests of knowledge—without addressing the real-world questions encountered by GPs.

Consequently, constructing a test set tailored to authentic scenarios in general practice is of critical importance. Such a test set not only provides data more reflective of real-life applications—helping researchers and developers better understand and improve LLM performance—but also strengthens oversight of these systems. Ultimately, this ensures their safety and effectiveness, fostering healthier and more responsible development of LLMs in the medical and healthcare sectors.

To address the aforementioned issues, we analyze the competencies required in the daily work of GPs. Drawing inspiration from the competency model for general practice, we design an evaluation framework for LLMs. Based on this target evaluation framework, we construct the necessary dataset as shown in Fig. 1, which consists of three components: the *MCQ Test Set*, the *Clinical Case Test Set*, and the AI Patient Test. The objective of the dataset is to comprehensively assess the required competencies. The MCQ Test Set evaluates the relevant medical knowledge of LLMs. The *Clinical Case Test Set* includes medical records from real-world settings, aiming to assess clinical capabilities in detail. The AI Patient Test is designed to determine whether LLMs can function like a real GP when interacting with a patient. The dataset is extensively annotated by experts and serves as a foundation for evaluating and analyzing state-of-the-art LLMs. Our main contributions are summarized as follows:

- **A novel evaluation framework for LLMs.** We borrow the core idea from the competency model for general practice and propose a novel framework to evaluate LLMs. The propped framework contains 6 primary indicators and 16 secondary indicators. It covers the main abilities needed in the daily work of GPs and provides a way to evaluate the performance of LLMs in the daily work of GPs comprehensively.
- **An fine-grained annotated dataset for real applications.** Based on the proposed general practitioner competence framework, we develop a Chinese general practice benchmark (GPBench), which can not only be used to evaluate the clinical diagnosis and treatment support capabilities of LLMs but also help identify weaknesses in the theoretical accumulation of models.
- **Evaluation of SOTA LLMs and performance analysis.** We conducted systematic evaluations of the current mainstream LLMs on the developed GPBench. The experiment results show that although these models show high accuracy in specific structured tasks, their comprehensive performance is systematically lacking in the clinical ability requirements of general practitioners in real medical scenarios. Combining with the proposed general practitioner competence framework, we found that

the LLMs' performance in the clinical diagnosis and treatment decision-making link lags significantly. The analysis reveals the cognitive gaps and reasoning limitations of current LLMs while providing priority directions for subsequent optimization of LLMs, which will help apply LLMs in real clinical settings.

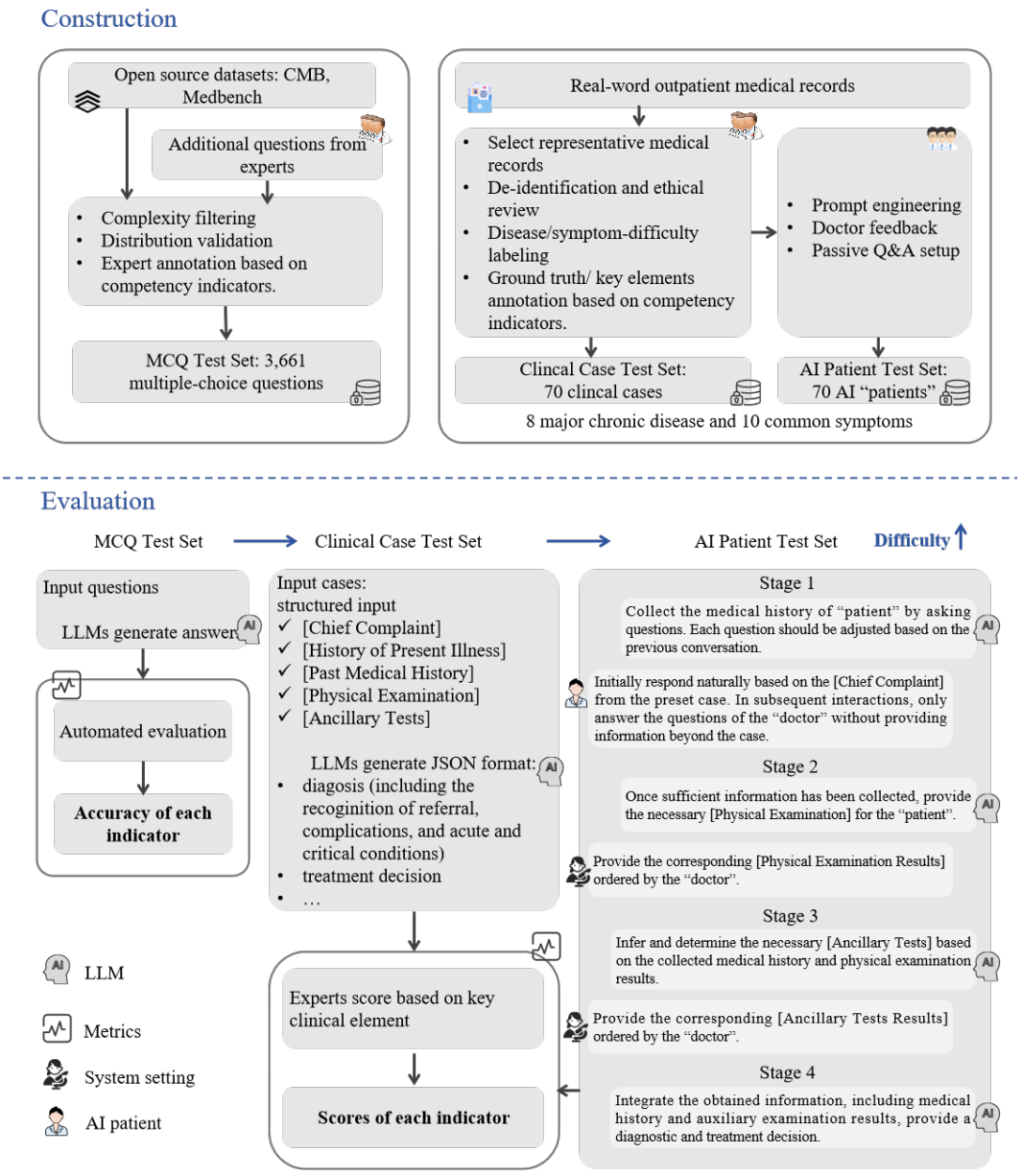

Figure 1: An overview of GPBench. For dataset construction, we compile multiple-choice questions from open-source datasets and real outpatient medical records from Tertiary A-grade hospitals to create three test sets: *MCQ Test Set*, *Clinical Case Test Set*, and *AI Patient Test*. Ground truth and scoring criteria for each case are annotated in details by experts. For evaluation, the GPBench framework innovatively employs the classic competency model of general practitioners to assess LLMs. Accuracy is used as the evaluation metric for the first test set, while for the other two test sets, experts grade responses based on the annotated scoring criteria.

2 Related Work

2.1 Large Language Models in Healthcare

Recent research on the application of LLMs in healthcare has made significant progress. There are mainly two types of LLMs in the healthcare area: general models and medical-special models. The general models are designed for wide applications and healthcare is just one of the target application areas. However, the medical-special models are designed only for the medical area. They are usually obtained by continual pre-training or finetuning the general model with medical corpus. The core challenge for fine-tuning is to preserve the general knowledge and semantic understanding ability while injecting medical knowledge and improving the medical reasoning ability.

2.1.1 General-purpose model

General models, like GPT-4 (OpenAI) [3], PaLM 2 (Google) [13], LLaMA family (Meta) [14, 15] and DeepSeek family (DeepSeek) [16, 17], have obtained extensive knowledge and strong semantic understanding capabilities through large-scale pre-training. These general-purpose models can be applied to the medical areas by setting proper prompts. Researchers have conducted tests on the performance of LLMs in medical scenarios. For example, in [7], researchers found that GPT-4 could exceed the passing score on the USMLE by over 20 points and outperformed GPT-3.5 as well as models specifically fine-tuned on medical knowledge (such as Med-PaLM, a prompt-tuned version of Flan-PaLM 540B). In [18], the AKT exam questions were used for testing, but the model did not achieve a passing score. In [19], it was found that general-purpose models require optimized prompts to achieve good performance. In [20], GPT-4 was tested, and the results showed that while it did not improve human-AI collaborative work, its performance significantly improved when working independently.

2.1.2 Medical-special Model

Usually, general-purpose models can be further improved with specially designed medical data. Thus, fine-tuning general LLMs is also widely explored by researchers in this field.

Data Source. Regarding data, some models use article-type corpora to inject medical-related knowledge. For example, HuatuoGPT-II [21] filters medical corpora from general datasets. BioMedLM [22], MEDITRON [23], BioMistral [24], and Clinical Came [25] utilize PubMed abstracts and full-text articles. MEDITRON [23] also incorporates medical guideline data. Such data sources typically contain large volumes of information, aiming to supplement LLMs' deficiencies in medical knowledge.

Other models use QA-type datasets. For instance, Med-PaLM [1], Med-PaLM2 [2], Med42-v2[26], and HuatuoGPT-o1[27] construct QA pairs extracted from training datasets. This type of data generally helps enhance the model's reasoning ability for specific medical questions. Med42 [28] leverages expert discussions, patient inquiries, and expert responses from medical forums such as the Stack Exchange network. ChatDoctor [29], Med42 [28], ClinicalGPT [30], and HuatuoGPT [31] use doctor-patient conversations and expert discussions extracted from medical forums and other sources. This type of data typically improves the model's conversational ability. Clinical Came [25] utilizes shared data from ShareGPT.

Some models use medical knowledge graphs for fine-tuning, such as ClinicalGPT [30], BenTsao [32], and Zhongjing [33]. ClinicalGPT [30] also fine-tunes using exam questions and electronic medical records. Other models leverage distilled data. For example, HuatuoGPT [31] incorporates medical instruction data and doctor-patient conversation data generated by ChatGPT.

Fine-Tuning Method. Regarding fine-tuning methods, a few models adopt continuous pre-training, such as MEDITRON [23] and BioMedLM [22]. However, as open-source model knowledge continues to improve, the space for enhancing general models' knowledge through continuous pre-training has diminished. Therefore, most models adopt supervised fine-tuning (SFT) to enhance reasoning abilities.

Some models further improve reasoning capabilities using reinforcement learning (RL), such as ClinicalGPT [30], Med42-v2[26], HuatuoGPT [31], HuatuoGPT-o1[27], and Zhongjing [33]. Among them, Zhongjing [33], HuatuoGPT-o1[27], and ClinicalGPT [30] adopt the PPO algorithm, while Med42-v2[26] uses the DPO algorithm. Generally, RL helps models generate outputs more aligned with human preferences.

Model Selection. Regarding model selection, early models used BLOOM due to the limited availability of open-source models at that time. Examples include ClinicalGPT [30] and HuatuoGPT [31]. As open-source models developed, the LLaMA series became a popular choice due to its strong performance. Models such as ChatDoctor [29], Med42 [28], Med42-v2[26], HuatuoGPT-o1[27], BenTsao [32], Clinical Came [25], MEDITRON [23], and MedAlpaca[34] have all adopted LLaMA-based models.

Other models have chosen different open-source architectures. For example, HuatuoGPT-II[21] is based on Baichuan2, BioMistral [24] is based on Mistral, and BianQue [35] is based on ChatGLM. Google's proprietary medical-special model, [36] and Med-PaLM2 [2], are based on PaLM 2 [13].

Performance Improvements. After fine-tuning, these models claim to achieve better performance than their base models, particularly on benchmark datasets that require reasoning capabilities. For example, BioMedLM [22] and Med-PaLM2 [2] have improved long-text QA capabilities after fine-tuning. BianQue [35] has enhanced proactive questioning ability. HuatuoGPT [31] has significantly improved multi-turn dialogue-based diagnostic capabilities. BenTsao [32] has significantly enhanced safety.

2.2 Evaluations and Benchmarks in Healthcare

To measure the development of LLMs in the healthcare field, researchers have developed various evaluation benchmarks to assess models' grasp of medical knowledge and clinical decision-making abilities from different perspectives. Basic medical knowledge is an important criterion for evaluating the medical capabilities of LLMs, often assessed through question-answer and multiple-choice formats to construct clinical datasets.

2.2.1 Multiple-choice question dataset

MedQA [37] and CMExam [38] are designed to evaluate models' basic medical knowledge by using multiple-choice question banks from medical exams. The former is derived from the United States Medical Licensing Examination (USMLE), while the latter is based on the China National Medical Licensing Examination (CNMLE) and the National Medical Examination Center (NMLEC). PubMedQA [39] gathers research questions from PubMed title and abstract. Each question can be answered yes/no/maybe. MedMCQA [40] contains 6150 test questions drawn from AIIMS PG and NEET PG questions found on the web and in books. each question having four multiple-choice options. HeadQA[4] [41] is a dataset composed of multi-choice questions, and the questions come from exams to access a specialized position in the Spanish healthcare system. MLEC-QA [42] collected questions from the National Medical Licensing Examination in China (NMLEC), which is designed to evaluate professional knowledge and skills for those who want to be medical practitioners in China. There are $136,236$ questions in MLEC-QA, and each question contains five candidate options with one correct/best option and four incorrect or partially correct options. And 10% are selected as the test set. CMB-test [5] [43] contains 11200 questions in total multiple-choice questions, most of them are examination from a well-known website. MedConceptsQA [44] comprises over $800,000$ questions and answers covering medical concepts, including ICD10 and ICD9 diagnoses codes, ICD9-PROC and ICD10- PROC procedures codes, and ATC drug codes. Each question contains a single medical code and four optional answers. MedConceptsQA is to evaluate whether LLMs can understand medical codes.

It should be noted that multiple-choice and question-answer formats significantly differ from real clinical processes, necessitating a more comprehensive evaluation method to reflect LLMs' performance in actual clinical settings.

2.2.2 Free response dataset

Huatuo-26M-test [45] contains $6,000$ samples of question-answer pairs in Chinese. However, the datasets are created from the web and not verified by experts. Thus it is not suitable for clinch settings. For the question-answer dataset, the answers are not verified by expert. It is not suitable for the clinch settings. For now, there is no suitable free response dataset for the clinch GP settings.

A set of challenging real-world cases from the New England Journal of Medicine (NEJM) case reports was used in [36] to evaluate the performance of LLMs on the differential diagnosis task. NEJM regularly publish patient cases which is described and then an expert physician is asked to provide a differential diagnosis and a final diagnosis and diagnostic reasoning, based only on the patient's provided medical history and preliminary test results. The data needs a license from NEJM to be used.

MedBench [6] [46], is a comprehensive and standardized benchmarking system for Chinese medical LLM. It assembled 8 existing benchmarks and constructed 12 datasets from past examination papers of medical schools, specialized medical textbooks, and real clinical case histories. MedBench contains $300,901$ questions, covering 43 clinical specialties and performs multi-facet evaluation on medical LLM. MedBench supports evaluation for both structured text and free tests. Nine datasets are evaluated on the free text, and the rest 11 datasets are designed for evaluation for multi-choice questions, structured text, and closed text. The benchmark for differential diagnosis is based on multiple-choice questions, and the benchmark for treatment is also simplified. Thus, the differential diagnosis and treatment dataset is not suitable for evaluation for real settings.

RareBench [47] is a benchmark to evaluate the performance of LLMs as rare disease specialists, and it contains 4 tasks: Phenotype Extraction from Electronic Health Records, Screening for Specific Rare Diseases, Comparison

[4] `https://aghie.github.io/head-qa/`

[5] `https://github.com/FreedomIntelligence/CMB`

[6] `https://medbench.opencompass.org.cn`

Analysis of Common and Rare Diseases, Differential Diagnosis among Universal Rare Diseases. The focus of RareBench is the differential diagnosis of rare diseases.

CLIBENCH [48], developed using the MIMIC-IV dataset, is designed to assess the capabilities of LLMs in clinical decision-making. It defines four types of clinical decision tasks: **Discharge Diagnoses**: Generate International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) codes based on patient medical records. **Treatment Procedure Identification**: Recommend the initial treatment procedures required after hospital admission, utilizing ICD-10-PCS codes. **Laboratory Test Recommendation**: Suggest the initial laboratory tests necessary after admission, using Logical Observation Identifiers Names and Codes (LOINC). **Medication Prescription Suggestion**: Propose the initial medications to be prescribed post-admission, following the Anatomical Therapeutic Chemical (ATC) classification system. The evaluation metrics include precision, recall, and F1 score, with multi-granularity assessments ranging from coarse-grained to fine-grained, tailored to the specific task requirements.

In [12], the authors evaluate LLMs on the diagnosis of four common abdominal pathologies, and they conclude that the LLMs are not ready for autonomous clinical decision-making. However, the above analysis is not designed for GP, and no competency models are involved. Thus, the performance of LLMs on the different dimensions for GP is not clear.

MultiMedQA was proposed in [1], and it is a combination of 7 medical question-answering datasets, including MedQA, MedMCQA, PubMedQA, LiveQA, MedicationQA and MMLU clinical topics and HealthSearchQA. The format of MultiMedQA is a hybrid of multiple-choice QA and free text generation. The long answers to medical questions in the LiveQA, MedicationQA, and HealthSearchQA datasets are annotated by humans.

However, existing benchmarks exhibit several limitations that hinder their suitability for evaluating LLMs comprehensively.

2.3 Medical Capability Assessment Framework for LLMs

The current evaluation methods mainly utilize multiple-choice questions and open-ended questions. Multiple-choice questions primarily assess accuracy, while open-ended questions focus on similarity to ground truth. Multiple-choice questions can evaluate a large model's knowledge and reasoning abilities. However, these questions differ significantly from the real-world clinical problems that doctors face, leading some evaluations to adopt free-text assessments to better measure the model's capabilities.

Some evaluations also aim to cover multiple dimensions as comprehensively as possible to make the assessment more holistic. For example, MedBench evaluates five dimensions: Medical Language Understanding, Medical Language Generation, Medical Knowledge Question Answering, Complex Medical Reasoning, and Healthcare Safety and Ethics. MedBench includes both multiple-choice and open-ended questions. However, diagnosis-related questions are still in multiple-choice format, while treatment-related questions are open-ended, but the standard answers tend to be relatively simple.

MultiMedQA attempts to evaluate the quality of textual responses. It assesses whether the answers align with the consensus in the clinical and scientific community, whether they provide evidence supporting the conclusions, whether they are correct or contain omissions, and whether they exhibit biases regarding medical demographics. [12] conducted a human evaluation of large models in real-world scenarios based on four common abdominal pathologies. The assessment dimensions include whether the diagnosis and management decisions are correct.

The current evaluations do not focus on the daily work of doctors and have not established a reliable evaluation model for large models. Most evaluations concentrate on diagnosis and treatment, but the granularity of analysis for treatment plans is insufficient. However, a doctor's work also involves proactive medical history collection, prescribing tests and examinations, and managing patients' health.

In this paper, we draw inspiration from the competency assessment methods for general practitioners (GPs) to evaluate large models. We conduct fine-grained manual annotation and assessment for each competency point. This is also the first evaluation dataset with a reliable evaluation framework, fine-grained annotation, and assessment.

3 The Proposed GPBench

3.1 The Evaluation Framework

To evaluate the LLMs, we need to determine how the LLMs perform in the real work settings of GP. Thus, we can borrow ideas from GP evaluation.

In order to meet the need for evaluating large language models, we have drawn on classic general practitioner (GP) competency frameworks such as WONCA [49], ACGME [50], McClelland's competency research dictionary, the Iceberg Model [51], and the Onion Model [52]. Taking into account the actual working conditions of GPs

in China, we propose a general practice competency model specifically for assessing large language models, detailed in Table 1.

This model comprises six primary general practice competency indicators (basic medical knowledge, diagnosis, decision-making, health management, health economics, and medical ethics and humanities) and sixteen secondary indicators (basic medical knowledge, diagnosis and differential diagnosis capability, medical history taking proficiency, complication identification skill; acute and critical condition recognition, referral decision-making competence, formulation of optimal treatment plans, adverse drug reaction management, contraindications awareness, alternative therapy selection capacity, health education delivery expertise, tertiary prevention implementation, patient compliance improvement, family support provision ability, cost-effective care coordination, and humanistic care competency). Our model encompasses the primary competencies required in the daily work of general practitioners, which are also the capabilities expected of AI in the GP setting.

Table 1: The competency indicators and definitions for our proposed evaluation framework.

Primary Indicator	Secondary Indicator	Definition
Basic Medical Knowledge	Basic Medical Knowledge	Basic medical knowledge refers to the foundational scientific understanding that constitutes the core of medical practice, encompassing the structural, functional, biochemical and pathological mechanisms of the human body. It serves as the theoretical basis for clinical reasoning and evidence-based decision-making.
Diagnosis	Diagnosis and Differential Diagnosis Capability (including test interpretation)	Diagnosis refers to the physician's judgment of the disease based on the patient's medical history, symptoms, signs, and auxiliary examination results. Differential diagnosis is the process of distinguishing the patient's main complaint from other diseases and excluding the possibilities of other conditions.
	Medical History Taking Proficiency	The ability to collect patient information in a comprehensive and accurate manner during the diagnostic and therapeutic process, including the patient's physical symptoms, psychological, mental, social, and cultural factors, as well as family history.
	Complication Identification Skill	The ability to predict, diagnose, and manage potential complications that may arise during the patient's treatment process. Complications refer to other diseases or symptoms that arise during the course of a primary disease, either due to the disease itself or as a result of its treatment.
	Acute and Critical Condition Recognition	The ability to recognize conditions that occur suddenly, are critical in nature, and require urgent intervention, potentially leading to life-threatening situations.
	Referral Decision-making Competence	The ability to identify and recognize diseases that pose a threat to the patient's life and ensure timely and correct referral to appropriate specialists.
Decision-making	Formulation of Optimal Treatment Plan (with ancillary test selection)	The ability to select an appropriate treatment plan based on the patient's condition.

Continued on next page

Table 1 - Continued from previous page

Primary Indicator	Secondary Indicator	Definition
	Adverse Drug Reaction Management	Adverse Drug Reaction (ADR) Management Competence refers to the comprehensive capability of healthcare institutions and professionals to promptly identify, scientifically assess, effectively intervene, accurately document, and systematically prevent unintended harmful reactions following medication use in clinical practice. Its core objectives are to minimize drug-related risks, ensure patient safety, and improve clinical outcomes through optimized therapeutic strategies. This competence requires adherence to evidence-based medicine principles, clinical protocols, and regulatory requirements, emphasizing multidisciplinary collaboration and continuous quality improvement.
	Contraindications Awareness	The ability of healthcare professionals or relevant personnel to accurately identify, comprehend, and apply drug contraindications (i.e., situations where a specific medication is prohibited or not recommended). This includes knowledge of patient-specific factors (e.g., allergies, pregnancy, hepatic/renal impairment), drug-drug interactions, disease-specific contraindications, and the practical skill to avoid inappropriate medication use in clinical settings, thereby ensuring patient safety.
	Alternative Therapy Selection Capacity	Alternative Treatment Selection Capacity refers to the comprehensive ability of patients or healthcare providers to systematically evaluate, appropriately choose, and effectively apply non-conventional therapeutic approaches (e.g., acupuncture, herbal medicine, energy therapies) when standard treatments are unsuitable, impractical, or ineffective. This capacity encompasses scientific understanding of alternative therapies, risk-benefit analysis, resource accessibility, and individualized decision-making skills, while integrating cultural context, patient values, and ethical considerations.
Health Management	Health Education Delivery Expertise	The specific competency required to undertake individual and community health education responsibilities and effectively conduct health education activities.
	Tertiary Prevention Implementation	The ability to implement health measures from the perspectives of prevention, treatment, and rehabilitation to achieve "preventing diseases before they occur, preventing disease progression during illness, and preventing recurrence after illness."

Continued on next page

Table 1 - Continued from previous page

Primary Indicator	Secondary Indicator	Definition
	Patient Compliance Improvement	Patient Adherence Enhancement Capacity refers to the comprehensive ability of healthcare providers or care teams to systematically improve patients' active cooperation and sustained compliance with treatment plans, medication regimens, or lifestyle modifications through strategies such as education, behavioral interventions, and optimized communication. This capacity involves understanding individual patient needs, identifying barriers to adherence, designing tailored interventions, and building trust through technical support, psychological reinforcement, and integration of social resources to foster long-term patient-provider collaboration.
	Family Support Provision Ability	Family Support Provision Ability refers to the collective ability of family members to provide emotional, financial, and practical resources to one another. It encompasses the functionality of a family unit to address life challenges, meet individual needs, and enhance overall well-being. Key components include emotional bonding, financial assistance, daily caregiving, educational guidance, and crisis management, reflecting the family's cohesion, communication effectiveness, and resource allocation strategies.
Health Economics	Cost-effective Care Coordination	The ability to scientifically control healthcare costs without compromising the quality of medical services, striving to minimize healthcare service fees and resource utilization.
Medical Ethics and Humanities	Humanistic Care Competency	In medical and healthcare work, the communication between healthcare providers and patients regarding injury, illness, diagnosis, treatment, health, and related factors is primarily led by the healthcare provider. Through various comprehensive means of communication, the goal is to scientifically guide the treatment of the patient's condition, achieve mutual understanding, and establish a collaborative relationship based on trust, ultimately contributing to human health maintenance, medical development, and societal progress.

3.2 GPbench Overview and Characteristic

GPBench is a medical evaluation set designed to comprehensively assess the general practice competencies of LLMs in healthcare in China. Based on the specially designed evaluation framework, GPBench focuses on assessing model performance in key areas such as clinical decision support and doctor-patient communication. An overview is provided in Table 2. We can see that the proposed GPBench consist of three components: the *Multiple-choice Questions (MCQ) Test Set*, the *Clinical Case Test Set*, and the *AI Patient Test Set*, which are described as follows.

Table 2: Overview of the proposed benchmark.

Content	Test Set	Formart	Number
Open-Source Data and Experts Supplement	MCQ Test Set	objective questions	3,661
Outpatient medical records containing any of the following 8 major chronic diseases or 10 common symptoms: • 8 Major Chronic Diseases: – Hypertension (HTN) – Hyperlipidemia (HLD) – Coronary Artery Disease (CAD) – Chronic Kidney Disease (CKD) – Chronic Obstructive Pulmonary Disease (COPD) – Cerebrovascular Disease (CVD) – Diabetes Mellitus (DM) – Cancer (CA) • 10 Common Symptoms: – Fever – Edema – Emaciation – Chest Pain – Headache – Abdominal Pain – Hematochezia – Joint Pain – Jaundice – Cough	Clinical Case Test Set	open-ended generation	70
	AI Patient Test Set	open-ended interaction	70

- **MCQ Test Set.** The test set consists of 3,661 multiple-choice questions (MCQ) and aims to evaluate the foundational medical knowledge and theoretical competency of LLMs in general practice, focusing on verifying the completeness of the model's knowledge system.
- **Clinical Case Test Set.** Built on 70 real-world de-identified outpatient medical records, the test set adopts an open-ended analysis format, requiring LLMs to analyze complete outpatient cases and propose systematic diagnostic and treatment plans. It emphasizes the abilities of models to systematically analyze complex problems and identify blind spots within real-world clinical decision-making chains.
- **AI Patient Test Set.** Based on the same 70 real-world de-identified outpatient medical records, the test set simulates real-world consultations by constructing AI-driven "patients" that dynamically interact with the LLM. It evaluates the clinical responsiveness and decision-making capabilities of LLMs in a simulated outpatient setting.

The *Clinical Case Test Set* and *AI Patient Test Set* are both built on 70 real-world de-identified outpatient medical records, covering eight major chronic diseases and ten common symptoms encountered in general practice. The reason we chose these diseases is that they have a relatively high consultation frequency in our survey targeting general practitioners. Moreover, CKD has a high prevalence in China and shows a clear upward trend. CVD is one of the leading causes of death in China, while CA ranks second among non-communicable chronic diseases in terms of mortality. Therefore, we selected samples from these eight diseases as evaluation cases. The reason we chose these ten symptoms is that they frequently appear in clinical consultations with general practitioners.

Metrics. For the *MCQ Test Set*, we calculated the LLM's accuracy in answering questions. To more precisely analyze the knowledge ability in terms of competence dimensions, we labeled each question with the corresponding competence indicator. For the *Clinical Case Test*, experienced clinical experts—based on actual work requirements—provide detailed correct answers, e.g., differential diagnosis, treatment, health education, and whether further examinations are needed. They also tag each part of the answer with its corresponding competence indicator. These answers can be directly used as clinical treatment of the corresponding patient.

The experts then score the LLM's output according to these labeled correct answers. For the *AI Patient Test Set*, experienced clinical experts score the diagnosis and treatment plans (including physical examinations and ancillary tests) generated by the LLMs after conversations with AI patients. The scoring dimensions align with the evaluation framework we have proposed. The details of dataset construction, annotation, and metric computation of the three components are described in the following subsections.

3.3 Construction Details of GPbench

This section primarily outlines the construction process of GPBench, which comprises three meticulously designed test sets: *MCQ Test Set*, *Clinical Case Test Set*, and *AI Patient Test Set*. These test sets progressively increase in complexity, enabling a multi-tiered evaluation of LLMs. The *MCQ Test Set* evaluates the application of theoretical knowledge through objective questions, as detailed in Section 3.3.1. The *Clinical Case Test Set*, described in Section 3.3.2, assesses LLMs' capability to systematically analyze medical cases using rigorously screened and ethically reviewed real outpatient medical records. The *AI Patient Test Set*, presented in Section 3.3.3, leverages LLMs to generate AI patient agents, facilitating dynamic outpatient interaction simulations to evaluate LLMs' active inquiry and decision-making abilities in real clinical scenarios.

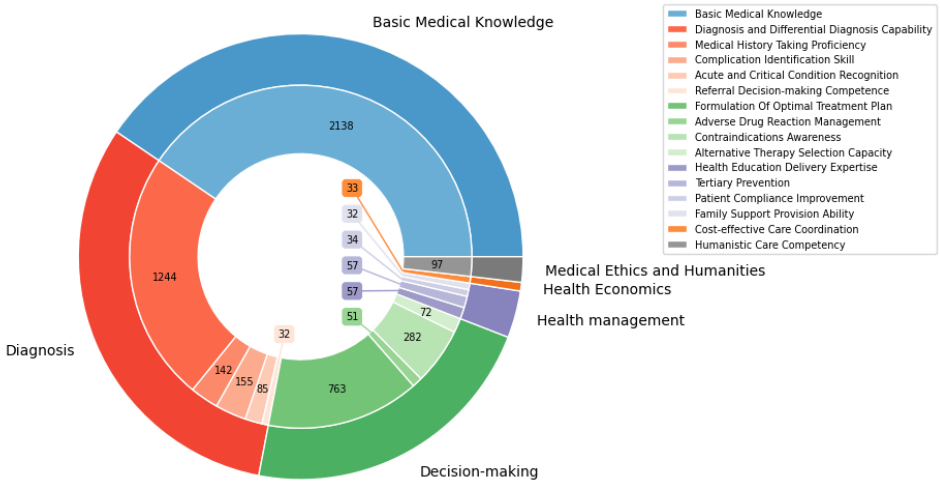

Figure 2: The data distribution on different medical competencies of *MCQ Test Set*.

3.3.1 MCQ Test Set

In the construction of the *MCQ Test Set*, we rigorously control the data sources, annotation process, and competency coverage to ensure its suitability for our setting. We select general practice-related questions from CMB [43] and MedBench [46] and annotate each question with the secondary indicators of the proposed evaluation framework. During this process, we observe that certain secondary indicators have a limited number of questions. To address this, clinical experts manually collect additional questions from other sources to enrich these underrepresented indicators. The final distribution and coverage of competency areas in the *MCQ Test Set* are illustrated in Figure 2. It can be seen that the number of questions corresponding to the indicators of Basic Medical Knowledge, Diagnosis, and Decision-Making is relatively high. This aligns with the general expectation that the daily work of general practitioners primarily involves applying basic medical knowledge to diagnose conditions and make clinical decisions.

Researchers [53] have proven that the order of options directly affects the evaluation results for multiple-choice questions. Thus, each question is tested with the options randomly sorted for multiple times. During the evaluation process, a question was considered correct only if the LLM answered it correctly in more than half of the repeated tests.

3.3.2 Clinical Case Test Set

Real-world outpatient medical records provide rich diagnostic and treatment information, including medical history, chief complaints, diagnoses, and treatment records. Our *Clinical Case Test Set* is developed based on 70 real-world outpatient medical records collected from multiple medical centers. These records have been meticulously curated and evaluated by expert physicians to ensure high standards of medical relevance, representativeness, and accuracy. All records have undergone strict de-identification and rigorous security and ethical reviews to safeguard patient privacy and data security. To comprehensively assess the ability of LLMs to

handle diverse clinical conditions, the test set is stratified into different difficulty levels for each disease category and symptom, thereby enhancing the robustness and reliability of the evaluation results.

For these anonymized medical records, experts with extensive clinical experience will annotate detailed correct answers based on practical work requirements. The answers include items such as differential diagnosis, management, health education, and whether further examinations are needed. These answers align with what a general practitioner would provide in a real work setting and can be directly used for the patient's clinical care. Furthermore, the general practice specialists also annotate the key points of each answer and outline the scoring criteria. Finally, they indicate which dimension of competency each part of the answer corresponds to. As for the output from the LLM, the experts score it according to the established scoring criteria. Considering the critical nature of medical decision-making and the strong dependence between diagnoses and treatment plans, we define the following scoring rules:

- 1. If the diagnosis is entirely incorrect, the treatment plan is assigned a score of 0 to ensure that erroneous diagnoses do not compromise the validity of model assessment.
- 2. If a diagnosis involves multiple diseases, the model's output is considered partially correct as long as it includes at least one accurate diagnosis. In such cases, the treatment plan is still evaluated, allowing for a more comprehensive assessment of the model's decision-making capabilities in complex clinical scenarios.

3.3.3 AI Patient Test Set

In the development of medical AI, it is extremely important to verify the clinical effectiveness of AI. Although models may pass preliminary validation through offline datasets, the complexity, and variability of real clinical environments differ significantly from standardized laboratory scenarios, which often lead to poor performance in practical applications. While real clinical testing is indispensable, it is costly and time-consuming, posing challenges for individual developers and small to medium-sized institutions. To address this problem, we develop the *AI Patient Test Set*, which simulates real clinical consultation scenarios to efficiently and cost-effectively validate AI performance. The interaction framework between the AI patient and the LLM is shown in Figure C.

We create AI patient agents for different diseases and symptoms, labeled with three difficulty levels: "easy", "medium", and "hard", based on labeled case data and prompt engineering. Specifically, we designed the prompts to make the responses of patient agents natural and conversational while strictly limiting their content to the information provided in the case, with detailed prompts available in Appendix A. During the prompt creation process, we incorporated feedback from outpatient doctors to ensure the responses aligned with clinical settings, making them as close as possible to real patient responses. Additionally, to further evaluate the model's follow-up questioning and clinical support abilities, we implement a passive $Q\&A$ setup. Each "patient" can only respond to questions asked by the "doctor" (LLM role) and will not proactively provide additional information, even if known. For example, the patient will not voluntarily mention their medical history or medication unless specifically asked. This mechanism effectively evaluates the ability of LLMs to ask appropriate follow-up questions to gather critical patient history information. The *AI Patient Test Set* focuses on assessing the clinical response capability and timely decision-making ability of models in real-world clinical scenarios. In the test set, we set a maximum number of inquiry rounds ($T = 10$) to evaluate whether the model can effectively extract key information within limited rounds, and then based on the obtained, make reasonable diagnostic and treatment recommendations. The scoring criteria for the test set are consistent with those set in the *Clinical Case Test Set*. Table 13 provides the key-element-based annotation results for a *Fever-medium* patient agent.

4 Results and Analysis

4.1 LLMs and Evaluation Metrics

To evaluate the state-of-the-art, we use GPbench to evaluate representative models from general and medical fields, including OpenAI's GPT series (GPT-4o, GPT-4-turbo, and GPT4-o1-preview), Google's Gemini-1.5-pro [54], Alibaba's Qwen series (Qwen2.5-7B-Instruct and Qwen2.5-72B-Instruct) [55], DeepSeek series (DeepsSeek-V3 [16] and DeepsSeek-R1 [17]), Claude-3.5-sonnet, and HuatuoGPT-o1-7B [56]. The base model of HuatuoGPT-o1-7B is Qwen2.5-7B-Instruct. We use the default parameters in its release website or the parameter settings recommended by the publisher. The information on the LLMs to be evaluated are shown in Table 3.

Since there are choice questions and open-ended questions in GPBench, we use a comprehensive set of indicators, shown in Table 1. Specifically, accuracy is used for choice answers, and expert-level manual annotation is used for open-ended generation, the details shown in Section 3.3.

Table 3: The LLMs evaluated in our experiments.

Model	Parameter	Type	
		Medical Specialist	Reasoning Model
GPT-4o [57]	-	No	No
GPT-4-turbo	-	No	No
o1-preview [58]	-	No	Yes
Gemini-1.5-pro [54]	-	No	No
Qwen2.5-7B-Instruct [55]	7B	No	No
Qwen2.5-72B-Instruct [55]	72B	No	No
Claude-3.5-sonnet	-	No	No
DeepSeek-V3 [16]	671B	No	No
DeepSeek-R1 [17]	671B	No	Yes
HuatuoGPT-o1-7B [56]	7B	Yes	Yes

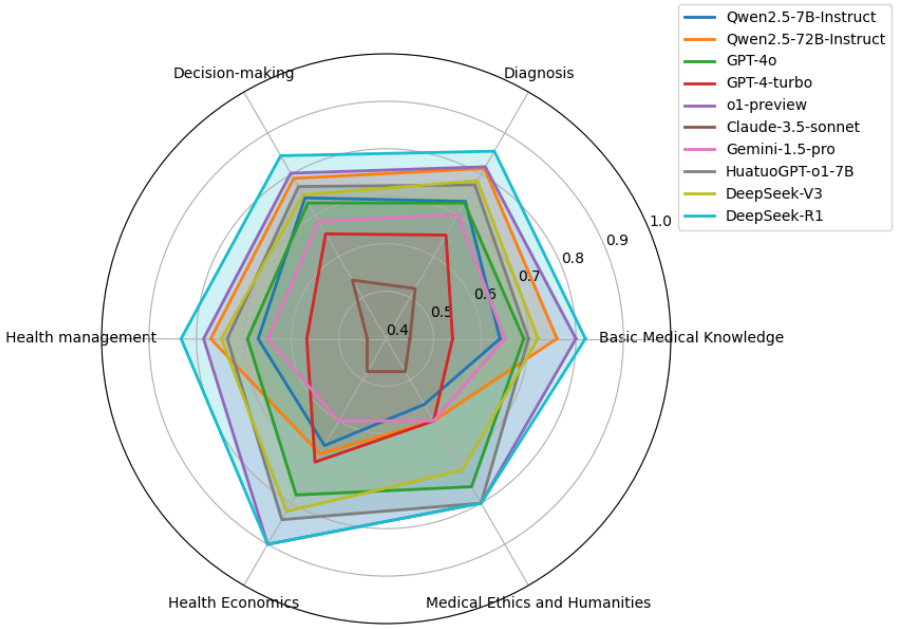

Figure 3: The performance of LLMs on the MCQ Test Set across the primary competency indicators.

4.2 Main Results

4.2.1 Evaluation Results on the MCQ Test Set

The performance of LLMs on the MCQ Test Set across the primary indicators is shown in Fig. 3. It can be observed that DeepSeek-R1 and o1-preview are the two best-performing models. This may be because reasoning ability plays a crucial role in answering questions in this test set, and these two models have been specifically enhanced in this aspect.

Table 4 presents the multi-dimensional accuracy performance of various LLMs in the *MCQ Test Set*. The results indicate that, except for o1-preview and DeepSeek-R1, most models exhibit relatively weak overall performance and significant shortcomings across multiple key dimensions. Although some models demonstrate competence in specific tasks, their overall stability and reliability remain insufficient to effectively support high-quality medical reasoning and clinical decision-making. In core dimensions such as foundational medical knowledge, treatment decision-making, and health economics, the majority of models perform suboptimally, highlighting the need for improvement in LLMs' medical knowledge coverage, clinical reasoning rigor, and adaptability to complex

decision-making. Specifically, 7/10 models scored below 75% in foundational medical knowledge accuracy, with HuatuoGPT-o1-7B achieving only 70.47%, indicating gaps in medical knowledge representation. In the four decision-making indicators (optimal treatment planning, adverse drug reaction management, contraindication comprehension, and alternative treatment strategies), even the relatively strong-performing o1-preview failed to exceed 75% accuracy in multiple metrics, while Claude-3.5-sonnet scored below 45% across all decision-related indicators. These findings underscore the current limitations of LLMs in clinical treatment knowledge integration and reasoning. In the health education dimension, models exhibited structural discrepancies in performance, particularly in the family support indicator, where accuracy was consistently lower than in other subcategories. This suggests that LLMs still struggle with fine-grained knowledge comprehension in health management. Furthermore, in cost-effective treatment planning, over half of the models achieved accuracy below 70%, reflecting gaps in LLMs' ability to integrate knowledge related to healthcare resource allocation and cost control.

o1-preview and DeepSeek-R1 demonstrated relatively superior performance, likely due to their stronger reasoning capabilities and more precise task comprehension, suggesting that models with advanced reasoning abilities have greater potential in general practice applications. Additionally, while HuatuoGPT-o1-7B excelled in medical ethics, humanistic care, and health economics, its performance in foundational medical knowledge, diagnosis, and treatment remained only at an upper-medium level. This disparity indicates that current LLM fine-tuning strategies have yet to achieve cross-dimensional optimization, making it challenging to ensure consistent advantages across different medical tasks.

4.2.2 Evaluation Results on the Clinical Case Test Set

In the *Clinical Case Test Set* evaluation, an expert panel used strictly annotated ground truth to grade different large language models (LLMs) on various dimensions of general practice competence using a 100-point scale. Tables 5 and 6 present the models' performance on common diseases and common symptoms, respectively, within the *Clinical Case Test Set*. The results indicate that LLMs exhibit systemic deficiencies in complex reasoning and decision-making tasks, particularly in the diagnosis and differential diagnosis, complications identification, and the formulation of optimal treatment plans.

Most models show significant shortcomings in the core aspects of clinical decision-making. In the diagnosis and differential diagnosis metric, over half of the models fail to construct a complete chain of diagnostic reasoning. The lowest-scoring model, HuatuoGPT-o1-7B, neglects key diseases in most cases and lacks a clear framework for managing chronic disease diagnosis and care. Regarding the formulation of optimal treatment plans, the majority of models do not effectively balance treatment efficacy, feasibility, and safety; in certain instances, some even provide medication recommendations that conflict with clinical guidelines. To delve deeper into the reasons behind poor performance in diagnosis and differential diagnosis, and in line with general practitioners' (GPs) clinical workflows, we introduced a new metric: Competency in Clinical Appropriateness Evaluation of Diagnostic Examinations. This metric gauges the model's capacity to analyze a patient's condition based on the chief complaint, medical history, and physical examination, and then determine which laboratory tests should be ordered. Nearly all models display systematic shortcomings in integrating patient histories to propose reasonable auxiliary examinations, either overlooking crucial tests or recommending superfluous ones.

Notably, DeepSeek-R1 stands out in both diagnosis and treatment, ranking first across six metrics. In terms of diagnosis, DeepSeek-R1 accurately stratifies and stages chronic diseases uses standardized diagnostic terminology, and provides comprehensive diagnoses. When devising treatment plans, DeepSeek-R1 adjusts strategies according to the patient's condition and medication contraindications, while emphasizing medication monitoring. This marked difference from other models highlights a potential optimization direction for future LLM applications in general medicine.

Moreover, despite explicit prompts, none of the LLMs proactively demonstrates humanistic care or cost-effective care coordination. This suggests that current LLMs remain limited to a simplified framework of diagnosis and treatment, lacking a broader understanding of the social attributes of healthcare services.

4.2.3 Evaluation Results on the AI Patient Test Set

In the *AI Patient Test Set* evaluation, an expert panel assigned hundred-point scores to various large models across multiple dimensions of general practice competence, based on strictly annotated ground truth. The results are presented in Tables 7 and 8. It can be seen that most models exhibit significant limitations in dynamic, interactive consultation scenarios. Comparing the performance of LLMs in the *AI Patient Test Set* and the *Clinical Case Test Set* reveals that overall multi-dimensional performance declines in the *AI Patient Test Set*, exposing systematic shortcomings in clinical inquiry practice. Notably, the reasoning-oriented model o1-preview stands out for its generally higher scores in multiple dimensions within the *AI Patient Test Set*. This may be due to the particular reinforcement of its reasoning process by simulating progressive information gathering in real consultations, enabling the model to iteratively refine its decision-making chain of reasoning during each round of dialogue and thus enhance its clinical reasoning abilities.

Table 4: The accuracies (%) of LLMs on the MCQ set across the secondary competency indicators.

Primary Indicator	Secondary Indicator	Qwen2.5-7B	Qwen2.5-72B	GPT-4o	GPT-4-turbo	o1-preview	Claude-3.5-sonnet	Gemini-1.5-pro	HuatuoGPT-o1-7B	DeepSeek-V3	DeepSeek-R1
Basic Medical Knowledge (15%)	Basic Medical Knowledge	64.00	76.47	68.78	54.50	79.88	45.28	64.83	70.47	71.86	82.44
Diagnosis (30%)	Diagnosis and Differential Diagnosis Capability	68.76	79.66	74.36	60.82	82.12	50.59	68.78	74.16	77.00	85.36
	Medical History Taking Proficiency	85.71	85.71	71.43	85.71	82.86	68.57	82.86	85.71	80.00	82.86
	Complication Identification Skill	77.68	82.61	82.32	63.77	86.38	48.99	69.28	81.74	82.03	87.54
	Acute and Critical Condition Recognition	65.95	80.83	72.96	62.66	83.12	47.93	69.38	72.95	78.97	89.27
	Referral Decision-making Competence	68.14	76.81	65.35	51.75	74.51	43.86	61.10	71.94	73.88	82.58
Decision-making (35%)	Formulation of Optimal Treatment Plan	72.83	78.11	77.36	72.08	86.04	62.64	76.23	76.89	76.60	84.53
	Adverse Drug Reaction Management	70.73	79.82	77.45	59.27	85.45	51.27	68.55	74.18	80.00	85.82
	Contraindications Awareness	72.50	80.00	65.00	62.50	75.00	47.50	50.00	70.00	67.50	85.00
	Alternative Therapy Selection Capacity	80.14	78.49	73.29	67.61	75.18	55.32	78.72	86.76	75.18	82.03
Health management (10%)	Health Education Delivery Expertise	66.86	77.71	69.71	58.29	74.86	44.00	63.43	75.43	74.86	80.57
	Tertiary Prevention Implementation	58.97	74.83	69.66	46.55	81.03	43.79	64.48	67.59	76.55	82.76
	Patient Compliance Improvement	64.12	74.12	63.53	54.71	72.35	42.65	60.00	66.67	71.18	79.12
	Family Support Provision Ability	77.86	81.07	72.86	67.14	85.71	45.36	73.21	83.57	76.07	89.64
Health Economics (5%)	Cost-effective Care Coordination	66.00	68.00	78.00	70.00	90.00	48.00	60.00	84.00	82.00	90.00
Medical Ethics and Humanities (5%)	Humanistic Care Competency	56.00	60.00	76.00	60.00	80.00	48.00	60.00	80.00	72.00	80.00

Table 5: Performance (scores) of LLMs for medical records containing the eight major chronic diseases in the Clinical Case Test Set.

Model	Diagnosis and Differential Diagnosis Capability (Weighted)	Competency in Clinical Appropriateness Evaluation of Diagnostic Examinations	Diagnosis and Differential Diagnosis Capability	Referral Decision-making Competence	Acute and Critical Condition Recognition	Complication Identification Skill	Formulation of Optimal Treatment Plan	Health Education Delivery Expertise
Qwen2.5-7B-Instruct	63.68	64.42	63.50	83.00	76.67	57.00	50.33	87.50
Qwen2.5-72B-Instruct	67.62	69.42	67.17	78.00	85.67	60.67	64.25	89.33
GPT-4o	65.30	61.83	66.17	84.67	86.67	74.17	61.67	86.83
GPT-4-turbo	61.68	62.42	61.50	83.00	88.67	70.67	57.33	83.33
Claude-3.5-sonnet	65.10	60.17	66.33	76.67	88.33	78.33	57.00	82.83
Gemini-1.5-pro	73.90	66.17	75.83	79.67	88.33	78.67	77.50	92.00
DeepSeek-V3	52.27	48.00	53.33	83.33	86.67	72.00	66.33	85.33
DeepSeek-R1	76.73	63.67	80.00	86.00	91.67	74.33	79.00	88.83
HuatuoGPT-o1-7B	45.17	42.50	45.83	78.33	85.00	68.67	62.00	90.00
o1-preview	66.47	67.00	66.33	89.67	90.00	76.67	63.75	89.67

Table 6: Performance (scores) of LLMs for medical records containing the ten common symptoms in the Clinical Case Test Set.

Model	Diagnosis and Differential Diagnosis Capability (Weighted)	Competency in Clinical Appropriateness Evaluation of Diagnostic Examinations	Diagnosis and Differential Diagnosis Capability	Referral Decision-making Competence	Acute and Critical Condition Recognition	Complication Identification Skill	Formulation of Optimal Treatment Plan	Health Education Delivery Expertise
Qwen2.5-7B-Instruct	64.18	67.88	63.25	70.25	84.00	65.75	44.18	78.00
Qwen2.5-72B-Instruct	69.25	74.75	67.88	71.25	90.00	60.13	57.50	86.50
GPT-4o	73.83	72.63	74.13	79.00	87.50	73.50	61.56	84.00
GPT-4-turbo	63.33	66.25	62.60	75.75	80.00	64.75	54.31	80.50
Claude-3.5-sonnet	71.15	70.25	71.38	78.25	82.50	69.00	59.06	77.00
Gemini-1.5-pro	74.18	63.88	76.75	78.25	82.50	68.00	59.13	84.00
DeepSeek-V3	68.80	62.00	70.50	67.00	81.25	67.25	61.50	83.00
DeepSeek-R1	79.68	85.88	78.13	75.23	96.25	78.25	76.75	87.25
HuatuoGPT-o1-7B	61.65	76.25	58.00	67.50	90.00	71.50	55.38	89.00
o1-preview	67.73	64.13	68.63	82.00	82.50	63.00	51.63	81.50

Table 7: Performance (scores) of LLMs for AI patients with the eight major chronic diseases in the AI Patient Test Set.

Model	Diagnosis and Differential Diagnosis Capability (Weighted)	Competency in Clinical Appropriateness Evaluation of Diagnostic Examinations	Diagnosis and Differential Diagnosis Capability	Referral Decision-making Competence	Acute and Critical Condition Recognition	Complication Identification Skill	Formulation of Optimal Treatment Plan	Health Education Delivery Expertise	Medical History Taking Proficiency
Qwen2.5-7B-Instruct	60.83	62.83	60.33	71.67	84.67	81.33	49.00	63.17	40.20
Qwen2.5-72B-Instruct	64.67	69.33	63.50	80.00	91.33	84.83	65.33	71.50	44.43
GPT-4o	67.47	70.00	66.83	75.83	94.67	84.00	58.83	73.33	48.13
GPT-4-turbo	59.90	68.17	57.83	67.17	93.00	79.83	51.67	69.17	40.17
Claude-3.5-sonnet	72.20	85.00	69.00	73.33	93.00	84.67	61.67	74.67	56.80
Gemini-1.5-pro	65.33	74.00	63.17	70.17	83.00	78.00	67.00	77.00	44.67
DeepSeek-V3	61.47	55.33	63.00	92.00	80.67	80.33	54.83	69.50	39.67
DeepSeek-R1	-	-	-	-	-	-	-	-	-
HuatuoGPT-o1-7B	41.57	45.17	40.67	43.33	59.00	47.00	26.00	41.03	38.03
o1-preview	72.13	68.67	73.00	97.33	93.00	86.00	70.33	90.00	55.00

In terms of humanistic care, cost-effective care coordination, and similar metrics, all LLMs show the same deficiencies observed in the *Clinical Case Test Set*.

Most LLMs demonstrate marked shortcomings in complex clinical reasoning and decision-making. In the "Medical History Taking Proficiency" metric, most models scored below 60, with the worst performer scoring only 32.13. This deficiency directly undermines the accuracy of subsequent diagnosis and treatment decisions. The models generally fail to gather adequate patient medical history, especially regarding personal history, which prevents them from offering personalized treatment plans and health education recommendations. In the "Competency in Clinical Appropriateness Evaluation of Diagnostic Examinations" metric, most models scored below 70, commonly missing key diagnostic items or suggesting redundant examinations.

It is worth noting that during the evaluation, DeepSeek-R1 showed a serious deficiency in following instructions. Within the standard evaluation process, DeepSeek-R1 frequently generated irrelevant and redundant answers unrelated to the current diagnostic stage, thus disrupting the evaluation process. Because it could not meet the basic evaluation requirements, DeepSeek-R1 was excluded from the *AI Patient Test Set* results system.

4.3 Analysis

This section primarily conducts a multi-dimensional analysis based on the test results of the *Clinical Cases Test*, with a focus on the proposed evaluation framework. Since none of the test results of LLMs include the indicator of cost-effective care coordination, this highlights the deficiency of LLMs in the dimension of health economics. This section presents a systematic and comprehensive evaluation from four perspectives: diagnosis, treatment decision-making, health management, and medical ethics.

4.3.1 Assessment of Diagnostic Capability

We perform an in-depth analysis of the performance of the five secondary indicators in the diagnostic dimension of the evaluated LLMs, uncovering their common deficiencies in the following aspects. The details are presented in Table 9.

Absence of Disease Classification and Staging System
Most models tend to oversimplify the management of diseases that require classification or risk stratification, lacking a dynamic modeling approach to disease progression. Consequently, the clinical diagnostic results generated by LLMs often lack hierarchy and depth. The average proportion of cases affected by this issue across all LLMs is 18.14%. For instance, in the *HTN-easy* case test, all evaluated LLMs fail to incorporate hypertension (HTN) classification or risk stratification, including Qwen2.5-72B-Instruct, o1-preview, Gemini-1.5-pro, and DeepSeek-R1. Notably, some models, such as GPT-4o, Claude-3.5-sonnet, and Gemini-1.5-pro, even exhibit

Table 8: Performance (scores) of LLMs for AI patients with the ten common symptoms in the AI Patient Test Set.

Model	Diagnosis and Differential Diagnosis Capability (Weighted)	Competency in Clinical Appropriateness Evaluation of Diagnostic Examinations	Diagnosis and Differential Diagnosis Capability	Referral Decision-making Competence	Acute and Critical Condition Recognition	Complication Identification Skill	Formulation of Optimal Treatment Plan	Health Education Delivery Expertise	Medical History Taking Proficiency
Qwen2.5-7B-Instruct	70.10	72.50	69.50	67.50	85.25	92.25	54.56	64.00	34.83
Qwen2.5-72B-Instruct	75.43	73.13	76.00	78.25	88.25	89.25	62.31	72.50	43.70
GPT-4o	80.93	80.13	81.13	77.25	91.50	94.00	63.88	71.63	53.30
GPT-4-turbo	72.98	75.38	72.38	65.75	86.75	91.75	59.00	68.13	43.48
Claude-3.5-sonnet	73.73	75.63	73.25	68.50	90.75	91.00	63.25	73.50	51.33
Gemini-1.5-pro	79.75	75.75	80.75	71.75	91.25	95.25	66.88	75.63	48.93
DeepSeek-V3	78.79	89.36	76.15	83.33	93.85	91.15	52.56	72.56	35.26
DeepSeek-R1	-	-	-	-	-	-	-	-	-
HuatuoGPT-o1-7B	65.13	75.38	62.56	33.59	76.15	73.59	40.00	37.18	35.05
o1-preview	77.69	72.56	78.97	96.67	99.23	92.82	62.82	93.59	55.23

classification errors, where systolic/diastolic blood pressure grading is misaligned. Similarly, in the case tests for coronary artery disease (CAD), most models omit cardiac function classification or lesion localization in their diagnostic results. Specifically, in the *CAD-easy* case test, $7/10$ models only provide a general diagnosis of "coronary artery disease/stable angina" without specifying cardiac function classification. The affected models include Qwen2.5-72B-Instruct, GPT-4o, and o1-preview. In the case tests for chronic kidney disease (CKD), most models exhibit inaccuracies in staging, with some even providing incorrect stage classifications. For example, in the *CKD-easy* case test, Qwen2.5-72B-Instruct and GPT-4o misdiagnose CKD stage 5 as CKD stage 3, while o1-preview only provides a generic diagnosis of CKD without specifying the disease stage. These findings indicate that existing LLMs generally struggle with the grading and staging of complex diseases. This limitation is likely due to insufficient exposure to relevant cases in training data and the incomplete construction of medical knowledge graphs.

Hallucination on Grading and Staging
LLMs arbitrarily grade and stage diseases without clinical guidelines or evidence-based medical support, leading to a significant issue of fabricating clinical staging during diagnosis. The average proportion of cases in which LLMs exhibit this issue is 22.57%. For certain diseases, such as pneumococcal pneumonia, diabetes, gouty arthritis, osteoarthritis, pneumothorax, and acute appendicitis, which lack well-defined clinical grading and staging standards, various models, including DeepSeek-R1, have generated fabricated grading and staging to different extents. Specifically, in the *Fever-easy* case test, the diagnostic results of Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct, and Claude-3.5-sonnet are pneumonia stage II," pneumonia (moderate grade 2)," and pneumonia (moderate)," respectively. In the *joint pain-medium* case test, Qwen2.5-72B-Instruct and GPT-4o provide the diagnoses of bilateral knee osteoarthritis stage 2" and knee osteoarthritis grade 2," respectively. Meanwhile, o1-preview and Gemini-1.5-pro both diagnose knee osteoarthritis exacerbation phase (Grade II)," and DeepSeek-R1 provides the diagnosis of "bilateral knee osteoarthritis (Grade 3)."

Blind Spots in Complication Recognition
Current LLMs generally exhibit insufficient awareness in systematically screening for comorbidities and complications, especially in the scenarios of metabolic and cardiovascular diseases, where key diagnostic elements are frequently omitted. This significantly impacts the comprehensiveness and accuracy of diagnoses. The average proportion of all LLMs in all cases that have this issue is 30.00%. The issue is primarily manifested in two aspects: **1) Failure to Fully Identify Complications.** LLMs fail to fully identify complications, which are consequently excluded from treatment plans, thereby reducing the effectiveness of treatment. For example, in the *HLD-hard* case test, almost all models, including GPT-4o, o1-preview, and DeepSeek-V3, failed to diagnose additional comorbidities or complications. Notably, even the highest-scoring models, Qwen2.5-72B-Instruct (score 70) and Gemini-1.5-pro (score 70), do not give a diagnosis of severe fatty liver and HLD. This omission may further affect long-term patient health management and prognostic assessment. **2) Incorrect assessment of complications.** Some LLMs may incorrectly diagnose complications, leading to erroneous treatment directions

Table 9: The percentages of observed deficiencies in the diagnostic process exhibited by LLMs.

Model	Absence of Classification and Staging System	Hallucination on Grading and Staging	Blind Spots in Complication Recognition	Deficiency in Acute and Severe Condition Assessment	Lack of Rare Disease Diagnosis
Qwen-2.5-7B-Instruct	28.57	22.86	32.86	4.29	4.29
Qwen-2.5-72B-Instruct	20.00	60.00	18.57	1.43	1.43
GPT-4o	12.86	15.71	37.14	2.86	1.43
GPT-4-turbo	18.57	8.57	35.71	4.29	8.57
Claude-3.5-sonnet	15.71	24.29	37.14	4.29	2.86
Gemini-1.5-pro	15.71	20.00	27.14	2.86	0.00
DeepSeek-V3	21.43	21.43	31.43	4.29	4.29
DeepSeek-R1	10.00	25.71	11.43	5.71	1.43
HuatuoGPT-o1-7B	17.14	1.43	44.29	2.86	8.57
o1-preview	21.43	25.71	24.29	2.86	4.29

or unnecessary interventions. For instance, in the *HLD-easy* case test, Qwen2.5-72B-Instruct and Gemini-1.5-pro incorrectly diagnosed diabetic nephropathy(DN), although the patient does not meet the diagnostic criteria for DN. Moreover, despite diagnosing DN, the models do not address its treatment in subsequent therapeutic plans. In real clinical scenarios, the above misdiagnoses may lead to the following ethical risks.

- **Patient psychological and health risks.** Misdiagnosis can cause patients to experience anxiety, fear, or depression, potentially triggering unnecessary psychological crises. Additionally, patients may undergo unnecessary invasive examinations (e.g., renal biopsy) or pharmacological treatments due to misdiagnosis, increasing the risk of adverse effects.
- **Waste of medical resources and crisis of trust.** Extra examinations, treatments, and follow-ups consume medical resources. Moreover, the trust of patients in AI-driven diagnostics and medical professionals may decline, leading to decreased adherence or even rejection of reasonable treatment plans.

Deficiency in Acute and Severe Condition Assessment
LLMs show significant shortcomings in the identification of clinical emergencies, particularly in the recognition of time-sensitive acute and life-threatening critical conditions. These models fail to accurately capture key characteristic indicators, which results in clinical recommendations that do not match the actual severity of the disease, potentially leading to adverse outcomes for patients. The average proportion of all LLMs in all cases that have this issue is 3.57%. This issue is mainly manifested in the following two aspects. **1) Failure to correctly identify and diagnose critical conditions.** The models fail to correctly identify and diagnose critical conditions, resulting in a lack of timely and effective intervention. For instance, in the *HTN-hard* case test, some models (e.g., Qwen2.5-7B-Instruct, GPT-4-turbo, Claude-3.5-sonnet) fail to identify hypertensive emergencies. In the *HLD-hard* case test, some models (e.g., Qwen2.5-7B-Instruct, GPT-4o) only diagnose acute pancreatitis but fail to identify severe pancreatitis. In such cases, the absence of targeted interventions may delay optimal treatment, increasing the risk of adverse outcomes. **2) Inconsistency between diagnosis and treatment.** Despite diagnostic results that fail to indicate critical conditions, some models adapt treatment recommendations for such conditions. For example, in the *HTN-hard* case test, some models (e.g., Qwen2.5-72B-Instruct, GPT-4o, o1-preview, Gemini-1.5-pro) do not recognize hypertensive emergencies in their diagnostic results, yet the subsequent treatment recommendations manage the condition as a hypertensive emergency. This reflects a disconnect between diagnostic results and the actual clinical condition. Although such management may alleviate acute symptoms to some extent, it fails to provide clear guidance for subsequent treatment through accurate diagnosis, potentially causing unnecessary resource waste or communication barriers in doctor-patient communication.

Lack of Rare Disease Diagnosis
LLMs exhibit limitations in diagnosing rare diseases, primarily due to the extreme scarcity of data samples, which makes it difficult for LLMs to acquire sufficient high-quality cases during training. Additionally, rare diseases often involve complex pathological mechanisms and heterogeneous clinical manifestations, increasing the likelihood of errors in reasoning and generalization. These limitations can lead to serious consequences, such as misdiagnosis or missed diagnoses, potentially depriving patients of optimal treatment opportunities and even exacerbating their conditions. The average proportion of all LLMs in all cases that have this issue is 3.71%. For instance, in the *Fever-medium* case test, $9/10$ LLMs, including DeepSeek-R1, failed to comprehensively assess the clinical presentation of patients and provide an accurate diagnosis for the rare disease scrub typhus.

4.3.2 Assessment of Treatment Recommendation

We conducted an in-depth analysis of the performance of LLMs evaluated on four secondary indicators in the decision-making dimension, revealing the common deficiencies of LLMs in the following aspects. The percentages of observed deficiencies in treatment recommendations are shown in Table 10.

Lack of Treatment Goals
In the treatment recommendations for chronic diseases such as hypertension and diabetes, there is a widespread lack of treatment goals. Specifically, almost all models, such as Qwen2.5-72B-Instruct, o1-preview, and Claude-3.5-sonnet, fail to explicitly set control targets for blood pressure or blood glucose when formulating treatment plans. Although some models, such as Gemini-1.5-pro and o1-preview, provide control targets, these targets are not set following the latest guidelines, which may adversely affect the assessment of treatment effectiveness and patients' long-term prognosis. For example, in the *HTN-easy* case test, all LLMs (including GPT-4o, o1-preview, Claude-3.5-sonnet, DeepSeek-R1, etc.) only suggest controlling blood pressure without specifying a concrete target value. In the *HTN-medium* case test, the o1-preview suggests a blood pressure control target of $130-140/80-90$ mmHg, which deviates from the latest guideline-recommended standard ($130/80$ mmHg). This inaccuracy and absence of target setting may lead to a lack of clear direction and evaluation criteria in the implementation of treatment plans, thereby affecting treatment outcomes and long-term health management of patients.

Pharmacotherapy Risks
Based on the evaluation results, we found that LLMs have issues such as omission of core drugs, neglect of drug interaction contraindications, lack of drug dosage guidance, and improper control of medication indications, which may lead to adverse drug reactions or treatment failure.

1) Omission of Core Drugs
Most models systematically omit core drugs in specific disease recommendations, which affects the completeness and effectiveness of treatment. For example, in the *CAD-easy2* case test, $3/10$ models fail to recommend drugs for heart rate control, such as beta-blockers (e.g., bisoprolol fumarate, metoprolol succinate, metoprolol tartrate). These drugs are one of the cornerstone medications for CAD, reducing myocardial oxygen consumption and heart rate, improving prognosis, and reducing mortality in post-myocardial infarction patients. The omission of the drugs may lead to increased myocardial oxygen consumption, triggering angina or myocardial infarction, and long-term high heart rate can cause ventricular remodeling, leading to poor prognosis and increased mortality. In the *HLD-hard* case test, $9/10$ models failed to include somatostatin and proton pump inhibitors (PPIs) in the treatment plan for severe acute pancreatitis. Somatostatin can inhibit pancreatic enzyme secretion and autodigestion, protecting pancreatic cells and reducing the risk of complications. PPIs can inhibit gastric acid secretion, preventing stress ulcers. The absence of these drugs may increase the burden on front-line treatment, raise the risk of complications such as gastrointestinal bleeding, prolong the disease course, and increase mortality in critically ill patients. In the *CKD-easy* case test, $6/10$ models do not recommend metabolic regulatory drugs such as calcium supplements and sodium bicarbonate to correct calcium and acid metabolism. The omission of these drugs can lead to calcium-phosphate metabolic disorders, renal osteodystrophy, metabolic acidosis, and worsening renal function, increasing the risk of cardiovascular events, fractures, and death.

2) Neglect of Drug Interaction Contraindications
Some models fail to comprehensively evaluate the overall safety of drug combinations, focusing only on the individual effects of each drug while ignoring the potential for increased side effects or adverse reactions when used together. This results in violations of the principles of drug combination therapy. For example, in the *HTN-easy* case test, Qwen2.5-7B-Instruct recommended the concurrent use of fibrate and statin drugs (fenofibrate and atorvastatin) for lipid-lowering, which may lead to an increased risk of overlapping drug side effects and adverse reactions, such as elevated risk of liver damage and increased risk of muscle-related adverse events. Similarly, in the *CAD-easy1* case test, Qwen2.5-72B-Instruct suggests the combined use of two beta-blockers (metoprolol and bisoprolol fumarate), which contravenes the basic principles of drug synergism. This could significantly increase the patient's risk of cardiac depression and negatively impact treatment efficacy.

3) Lack of Drug Dosage Guidance
Some models show insufficient or erroneous guidance on drug dosing, posing potential risks to the safety and efficacy of clinical medication use. For example, in the *HLD-hard* case test, GPT-4-turbo fails to specify the dosages for lipid-lowering drugs and antibiotics, leaving clinical practice without a reliable basis. In the same case, Qwen2.5-72B recommends a dose of Ulinastatin at $10,000\ U$, which significantly deviates from the $100,000\ U$ recommended by the guidelines. Moreover, in some cases, certain models even suggest dosing routes that do not match the acute phase of the disease. For example, in the *COPD -medium* case test, where the patient was experiencing an acute exacerbation of COPD (with worsening cough, sputum production, and elevated infection markers), GPT-4o recommends Amoxicillin-Clavulanate $500\ mg$, orally every 8 hour instead of intravenous antibiotic therapy. This could result in the patient's symptoms not being controlled in a timely manner.

4) Improper control of medication indications

Some models fail to adequately consider individual patient characteristics and specific disease requirements when providing treatment recommendations, resulting in improper control of medication indications. For example, in the *CKD-easy1* case test, some models, including Qwen2.5-7B-Instrcut, Qwen2.5-72B-Instrcut, and Claude-3.5-sonnet, do not account for the patient's renal function when selecting and adjusting antihyperglycemic drugs and then fail to recommend discontinuing metformin. Given that renal impairment affects the excretion of metformin, such a dosing regimen may lead to severe adverse reactions such as lactic acidosis. This indicates that LLMs fail to fully recognize the impact of renal function on drug metabolism and safety. Similarly, in the *HTN-medium* case test, the treatment plan from Qwen2.5-7B-Instruct do not consider the patient's documented adverse reactions to ACEI drugs (e.g., benazepril). The oversight of individual patient drug reactions may reduce treatment safety and increase the risk of recurrent adverse reactions. Furthermore, some models exhibit inappropriate use of antibiotics without clear indications. For instance, in the *CKD-easy1* case test, the patient lacks indications for antibiotic use, yet GPT-4-turbo recommends norfloxacin. Such unwarranted use of antibiotics increases the risk of antimicrobial misuse and imposes unnecessary economic burdens and risks of adverse drug reactions on patients.

Blind Spots in Non-Pharmacological Interventions

Based on the test results, we find that the LLMs' recommendations for non-pharmacological treatment plans, such as interventional therapies and surgical indications, are significantly underrepresented. However, in certain disease scenarios, the absence of non-pharmacological treatment options may directly impact the comprehensive treatment outcomes and prognosis of patients. For example, in the *CAD-medium* case test, the patient presents with acute myocardial infarction. However, some models, such as Qwen2.5-72B-Instruct, GPT-4o, and GPT-4-turbo, do not include critical treatment options such as emergency percutaneous transluminal coronary angioplasty (PTCA), stent implantation, or thrombolysis. Acute myocardial infarction is a severe complication of coronary heart disease, and timely interventional or thrombolytic therapy is crucial for restoring myocardial perfusion and reducing myocardial damage. The failure of LLMs to provide these key interventional treatment options may cause patients to miss the optimal treatment window and increase the extent of myocardial infarction and the risk of complications, thereby affecting long-term prognosis. In the *CKD-hard1* case test, the patient has progressed to CKD Stage 5 (eGFR $< 15ml/min/1.73m^2$). However, some models, such as Qwen2.5-7B-Instruct, Qwen2.5-72B-Instruct, GPT-4-turbo, and Gemini-1.5-pro, do not provide a dialysis plan tailored to the stage of disease progression. CKD Stage 5 represents end-stage renal disease, and dialysis is a critical measure to sustain the patient's life. The failure of LLMs to recommend dialysis may lead to further deterioration of renal function and even life-threatening conditions.

Insufficient Standardization in Critical Illness Management

In the management recommendations for critical and life-threatening conditions, some models fail to adhere to clinical standards, and even omit critical steps in certain disease scenarios, which could pose a fatal threat to patient safety in real clinical practice. Specifically, in the *DM-hard* case test, where the patient is in a state of Diabetic Ketoacidosis, Claude-3.5-Sonnet suggests potassium supplementation only in the presence of hypokalemia. This is in stark contrast to the clinical standard, which requires potassium replacement when urine output exceeds $30\ ml/h$ and serum potassium is below $5.2\ mmol/L$. In the same case, GPT-4-turbo fails to provide key treatment parameters such as insulin dosage and fluid resuscitation volume, which directly impacts the timely control of the condition. In the *CVD-hard* case test, where the patient is experiencing intracerebral hemorrhage, some models, such as Qwen2.5-7B-Instruct, GPT-4-turbo, and Claude-3.5-sonnet, recommend the use of oral antihypertensive medications. However, the *Chinese Guideline for the Diagnosis and Treatment of Intracerebral Hemorrhage (2019)* recommends the use of intravenous antihypertensive drugs during the acute phase of intracerebral hemorrhage to rapidly control blood pressure. Oral antihypertensive medications cannot promptly reduce blood pressure to the target range, which may lead to further exacerbation of intracerebral hemorrhage and increase the patient's risk of disability and mortality.

Differential Cross-Disease Capabilities

We obtain that LLMs exhibit significant differences in diagnostic and treatment scores across various diseases from Table 5. This phenomenon primarily stems from the varying demands placed on general practitioners by different diseases. In the case tests of CA, the models generally achieve high scores in the indicator of diagnosis, reaching 83.33 on average. This is because the primary task of general practitioners when dealing with patients in CA is to identify key indicators and refer them to oncology specialists. This relatively straightforward diagnostic and treatment task makes it easier for LLMs to meet basic clinical needs, thereby achieving higher scores. In contrast, the diagnosis and treatment management of chronic diseases place more complex and nuanced demands on models. For example, in the diagnosis of HTN and HLD, the average scores of the models are 58.33 and 74.08, respectively. This is because the management of HTN and HLD requires models to provide specific drug selections and dosage adjustments and address lifestyle modifications, complication management, and long-term follow-up. The deficiencies of LLMs in generating these nuanced treatment recommendations result in lower scores. This disparity highlights the models' current shortcomings in generating detailed treatment suggestions, especially in the management of complex chronic diseases, which means the LLMs' performance in these areas is still insufficient to meet the actual clinical needs.

Table 10: The percentages of observed deficiencies in treatment recommendations exhibited by the evaluated LLMs.

Models	Lack of Treatment Goals	Pharmacotherapy Risks				Blind Spots in Non-Pharmacological Interventions	Insufficient Standardization in Critical Illness Management
		Omission of Core Drugs	Neglect of Drug Interaction Contraindications	Lack of Drug Dosage Guidance	Improper control of medication indications		
Qwen-2.5-7B-Instruct	24.29	24.29	1.43	17.14	8.57	20.00	12.86
Qwen-2.5-72B-Instruct	22.86	25.71	1.43	10.00	4.29	12.86	11.43
GPT-4o	20.00	25.71	5.71	8.57	2.86	15.71	11.43
GPT-4-turbo	24.29	30.00	2.86	24.29	2.86	15.71	12.86
Claude-3.5-sonnet	25.71	34.29	2.86	17.14	2.86	14.29	11.43
Gemini-1.5-pro	15.71	22.86	1.43	14.29	0.00	15.71	10.00
DeepSeek-V3	25.71	28.57	7.14	7.14	1.43	17.14	11.43
DeepSeek-R1	17.14	7.14	2.86	2.86	1.43	17.14	2.86
HuatuoGPT-o1-7B	24.29	28.57	0.00	14.29	4.29	14.29	8.57
o1-preview	21.43	18.57	2.86	5.71	0.00	14.29	7.14

4.3.3 Assessment of Health education capability

Current LLMs present an overly principled approach when providing health education advice. Although the content generally meets basic scoring standards, it often remains at a broad, principle-based level, lacking concrete, actionable guidance that would effectively support clinical practice needs. For instance, in the case tests of HTN, most models merely suggest a "low-salt diet" without specifying precise daily sodium intake limits (e.g., $\leq 5g$), which significantly differs from the quantified targets outlined in the *Chinese Hypertension Prevention and Treatment Guidelines (2024)*. This ambiguity not only reduces patient compliance but may also weaken the practical effectiveness of interventions, ultimately impacting the overall quality of health management.

4.3.4 Assessment of Medical Ethical Risks

The tested LLMs pose the following ethical risks in the clinical diagnosis and treatment field.

Over-treatment Tendency
Some LLMs may generate unreasonable treatment recommendations due to algorithmic bias or limitations in training data. This is particularly evident in specific disease scenarios, where models may suggest excessive polypharmacy or treatments that do not align with the disease stage. This tendency exposes patients to unnecessary medical risks and increases their economic burden. For example, in the *HTN-easy* case test, Qwen2.5-7B-Instruct recommends the simultaneous use of beta-blockers and statins (fenofibrate and atorvastatin) for lipid-lowering, which is unnecessary for the current patient and inconsistent with clinical protocols. Although the recommendations provided by the models may have some reference value, over-reliance on model-generated results by physicians could lead to adverse outcomes in practice.

Risk of Misdiagnosis or Missed Diagnosis
When general practitioners overly rely on the judgments of LLMs in real clinical decision-making, there is a risk of misdiagnosis or missed diagnosis, especially if LLMs fail to accurately or comprehensively assess the severity of the patient's condition. For example, in the *HTN-hard* case test, the patient is experiencing a hypertensive emergency, but some models misdiagnose the condition as mild hypertension or an unrelated disorder, such as transient ischemic attack (e.g., Qwen2.5-7B-Instruct diagnoses hypertension, GPT-4-turbo diagnoses transient ischemic attack). This misjudgment may disrupt the doctors' clinical decision-making process. If the doctor relies on these inaccurate recommendations and fails to take timely and necessary actions, the patient may miss the optimal treatment window, potentially leading to severe medical incidents.

5 Conclusion

This study aims to evaluate the capability of existing large models in general practice clinical scenarios. To this end, a novel evaluation framework was proposed by drawing inspiration from the Competency Model for General Practice. Based on this framework, we constructed a dataset to assess the essential capabilities of

LLMs in general practice. The dataset consists of an *MCQ Test Set*, a *Clinical Case Test Set*, and an *AI Patient Test Set*. The dataset was manually annotated by experts to provide correct answers and scoring criteria. In particular, the samples in the *Clinical Case Test Set* and the *AI Patient Test Set* were designed according to the actual work requirements of general practitioners, encompassing all the outputs expected in real-world clinical practice. Therefore, compared with previous evaluation datasets, our dataset offers a more scientifically rigorous evaluation methodology and richer data details.

Based on the newly designed evaluation method and the annotated dataset, we assessed the performance of current state-of-the-art (SOTA) large models, including both general-purpose models and medical-specialized models. The experimental results indicate that both medical and general-purpose models still exhibit significant shortcomings in areas such as clinical decision support, diagnostic accuracy, and treatment recommendations. In particular, their performance remains insufficient when dealing with complex cases, providing personalized treatment suggestions, and reasoning through diagnostic and treatment processes. Although medical LLMs demonstrate a slight advantage over general-purpose LLMs in certain tasks, their overall performance still requires further improvement. To enhance the capability of large models in real-world general practice scenarios, we suggest that future research could focus on the following directions:

- Further improving the systematic medical knowledge and complex medical reasoning abilities of large models.
- Optimizing the generation of results to include more detailed information, especially the critical details required in the real-world work scenarios of general practitioners.

Through extensive collaboration and continuous innovation, we aspire to drive advancements in large models for general practice and contribute to the enhancement and optimization of clinical decision support systems.

References

[1] 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. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

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

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

[4] Eric Strong, Alicia DiGiammarino, Yingjie Weng, Andre Kumar, Poonam Hosamani, Jason Hom, and Jonathan H Chen. Chatbot vs medical student performance on free-response clinical reasoning examinations. *JAMA internal medicine*, 183(9):1028–1030, 2023.

[5] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312, 2023.

[6] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences (2076-3417)*, 11(14), 2021.

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

[8] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

[9] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200, 2023.

[10] Yanhui Zhang, Haolong Pei, Shihan Zhen, Qian Li, and Fengchao Liang. Chat generative pre-trained transformer (chatgpt) usage in healthcare. *Gastroenterology & Endoscopy*, 1(3):139–143, 2023.

[11] Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Parastoo Falakaflaki, Elena Kayayan, Guanyu Tao, Mahta Haghighat Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, et al. The potential and pitfalls of using a large language model such as chatgpt, gpt-4, or llama as a clinical assistant. *Journal of the American Medical Informatics Association*, 31(9):1884–1891, 2024.

[12] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

[13] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[14] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[18] Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, and Sachin Shah. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9(1):e46599, 2023.

[19] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[20] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.

[21] Junying Chen, Xidong Wang, Ke Ji, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Song Dingjie, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. HuatuoGPT-II, one-stage training for medical adaption of LLMs. In *First Conference on Language Modeling*, 2024.

[22] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

[23] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[24] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[25] Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *CoRR*, 2023.

[26] Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*, 2024.

[27] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

[28] Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*, 2024.

[29] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

[30] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

[31] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

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

[33] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376, 2024.

[34] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[35] Yirong Chen, Zhenyu Wang, Xiaofen Xing, Huimin Zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *ArXiv*, abs/2310.15896, 2023.

[36] Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *arXiv preprint arXiv:2312.00164*, 2023.

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

[38] Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452, 2023.

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

[40] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022.

[41] David Vilares and Carlos Gómez-Rodríguez. HEAD-QA: A healthcare dataset for complex reasoning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy, July 2019. Association for Computational Linguistics.

[42] Jing Li, Shangping Zhong, and Kaizhi Chen. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[43] Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*, 2023.

[44] Ofir Ben Shoham and Nadav Rappoport. Medconceptsqa: Open source medical concepts qa benchmark. *Computers in Biology and Medicine*, 182:109089, 2024.

[45] Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*, 2023.

[46] Mianxin Liu, Weiguo Hu, Jinru Ding, Jie Xu, Xiaoyang Li, Lifeng Zhu, Zhian Bai, Xiaoming Shi, Benyou Wang, Haitao Song, Pengfei Liu, Xiaofan Zhang, Shanshan Wang, Kang Li, Haofen Wang, Tong Ruan, Xuanjing Huang, Xin Sun, and Shaoting Zhang. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Mining and Analytics*, 2024.

[47] Xuanzhong Chen, Xiaohao Mao, Qihan Guo, Lun Wang, Shuyang Zhang, and Ting Chen. Rarebench: Can llms serve as rare diseases specialists? In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4850–4861, 2024.

[48] Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S Chang, and Wei Wang. Clibench: Multifaceted evaluation of large language models in clinical decisions on diagnoses, procedures, lab tests orders and prescriptions. *arXiv preprint arXiv:2406.09923*, 2024.

[49] WONCA. The european definition of general practice/family medicine, June 2012.

[50] Joseph E Scherger. Preparing the personal physician for practice (p4): essential skills for new family physicians and how residency programs may provide them. *The Journal of the American Board of Family Medicine*, 20(4):348–355, 2007.

[51] David C McClelland. Testing for competence rather than for" intelligence.". *American psychologist*, 28(1):1, 1973.

[52] Richard E Boyatzis. *The competent manager: A model for effective performance*. John Wiley & Sons, 1991.

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

[54] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[55] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[56] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024.

[57] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[58] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

A Samples from our annotated dataset

Table 11: An annotated example from the MCQ Test Set.

Field	Value
ID	1278
Question	女性，40岁。类风湿关节炎10年，长期服通用非甾体抗炎药，实验室检查血常规血红蛋白78g/L。关于该患者贫血的说法，错误的是
Options	A: 是小细胞低色素性贫血 B: 属于慢性病性贫血 C: 主要发生机制是铁利用障碍 D: 可能有缺铁因素参与 E: 常伴有血小板减少
Answer	E
Type	单项选择
Source	CMB-test-医师考试-中级职称-内科主治医师
Indicators	并发症的识别能力, 药物不良反应处理能力, 诊断与鉴别诊断

Table 12: An example of *Fever-medium* outpatient medical record.

Case	Symptom	Difficulty
患者李XX，47岁男性。 主诉：发热1周。 现病史：患者1周前爬山后开始出现发热，体温38°C，伴畏寒、寒战，伴咳嗽、咳痰，痰为白色粘痰，量多，伴肌肉酸痛，在当地卫生站就诊，查血常规提示血小板减低，经过治疗症状未见明显好转，反复发热，体温升至39°C，伴寒战，伴头晕、头痛，偶伴恶心，为进一步来诊。 既往史：有高血压病2级病史多年，现未服用降压药物，监测血压基本正常。 查体:T38°C,P120次/分，R28次/分，BP136/76 mmHg。神志清，左下肢见0.5cm*0.5cm皮肤破溃；双肺呼吸音稍弱，双下肺可闻及少量湿性音；心率120次/分，律齐，各瓣膜听诊区未闻及明显病理性杂音；腹平软，全腹部无压痛及反跳痛，肝脾肋下未触及；四肢肌力、肌张力正常；双下肢无浮肿。 辅助检查：血常规：WBC：8.96*10^9/L，NEUT%：50.6%，Hb：150.0g/L，PLT：48*10^9/L↓。DD2：13.06ug/ml↑，BNP：69.90pg/ml↑，hs-cTnI：0.033ng/ml，呼吸道三项、两对半、输血四项正常。生化：心肌酶正常，BUN：6.33mmol/L，CR：81.50umol/L，Ca：1.95mmol/L↓，hs-CRP：73.52mg/L↑，ALT：190.64U/L↑，AST：203.98U/L↑，r-GT：173.50U/L↑，ALB：28.50g/L↓，FER：>1025.0ng/ml↑。ECG: 1：窦性心律。2：肢体导联低电压。心脏彩超：轻微二尖瓣返流。轻度三尖瓣返流。左室收缩功能正常。腹部、泌尿系彩超：脾脏稍大。肝脏、胆囊、胆管、胰腺、声像图:未见明显异常。前列腺稍大。双侧肾脏大小正常范围。双侧输尿管:未见扩张。膀胱内未见异常。胸部CT报告：1.双肺多发散在炎症及纤维化灶，双侧胸腔少量积液，请复查。2.双侧胸膜增厚；拟右侧叶间胸膜包裹性积液，请结合临床。 诊断 (not seens by LLMs)： 1.恙虫病 2.脾大 3.肝功能不全 4.胸腔积液 5.低蛋白血症 6.继发性血小板减少 7.高血压病2级；高危组 治疗 (not seens by LLMs)： 一、一般治疗 患者应注意休息， 进食易于消化食物，注意口腔卫生， 维持水、电解质平衡， 减少并发症发生。高热可用物理降温，酌情使用解热药物，但慎用大量发汗的退热药。 二、病因治疗（药物选择其中一种即可） 1． 四环素类：多西环素推荐剂量：成人 0.1 g，每日 2 次，首剂加倍；抗感染疗程至少持续到体温复常后 3 d，临床症状改善并稳定，一般 7～10 d。重症患者可选择静脉滴注给药，成人第 1 天 0.2 g，分 1～2 次静滴，以后根据感染程度每日 0.1～0.2 g 静滴 2. 大环内酯类：罗红霉素成人 150 mg，每日 2 次，退热后 150 mg，每日 1 次。阿奇霉素成人 0.5 g，每日 1 次，退热后 0.25 g， 每日 1 次；克拉霉素成人 0.5 g/次，每 12 h 口服 1 次。以上药物疗程均为 7 10 d。 三、对症支持治疗：恙虫病合并呼吸、肝、 肾功能损害时，常采取抗炎、给氧、保肝降酶以及维持/替代肾功能等治疗。	Fever	Medium

Table 13: An example of annotated scoring criteria for the sample of *Fever-medium* medical record.

Indicator	Applicability	Scoring Criteria
Diagnosis and Differential Diagnosis Capability	Yes	诊断： 1.恙虫病（40分） 2.脾大（10分） 3.肝功能不全（10分） 4.胸腔积液（10分） 5.低蛋白血症（10分） 6.继发性血小板减少（10分） 7.高血压病2级；高危组（10分））
Referral Decision-making Competence	Yes	暂不需要转诊（100分）
Acute and Critical Condition Recognition	Yes	肝功能不全（25分）；胸腔积液（25分）；.低蛋白血症（25分）；继发性血小板减少（25分）
Complication Identification Skill	Yes	未涉及危急重症（100分）
Formulation Of Optimal Treatment Plan	Yes	治疗： 二、病因治疗（药物选择其中一种即可）（50分，若回答对药物名称，用法用量不正确扣30分） 1．四环素类：多西环素推荐剂量：成人 0.1 g，每日 2 次，首剂加倍；抗感染疗程至少持续到体温复常后 3 d，临床症状改善并稳定，一般 7～10 d。重症患者可选择静脉滴注给药，成人第 1 天 0.2 g，分 1～2 次静滴，以后根据感染程度每日 0.1～0.2 g 静滴 2. 大环内酯类：罗红霉素成人 150 mg，每日 2 次，退热后 150 mg，每日 1 次。阿奇霉素成人 0.5 g，每日 1 次，退热后 0.25 g，每日 1 次；克拉霉素成人 0.5 g/次，每 12 h 口服 1 次。以上药物疗程均为 7 10 d。 三、对症支持治疗（50分）：恙虫病合并呼吸、肝、肾功能损害时，常采取抗炎、给氧、保肝降酶以及维持/替代肾功能等治疗。
Further enhance examination and testing capabilities	Yes	N/A
Health Education Delivery Expertise	Yes	患者应注意休息，进食易于消化食物，注意口腔卫生，维持水、电解质平衡，减少并发症发生。高热可用物理降温，酌情使用解热药物，但慎用大量发汗的退热药（100分）
Cost-effective Care Coordination	No	N/A
Humanistic Care Competency	No	N/A

B Prompt for patient agents

Patient Agent prompt(CN)

角色：你是一名来医院看病的患者。
个人信息：[个人信息]
任务: 根据医生的问题，严格按照以下分类和行动指南进行回应，以模拟真实的医患对话。
目标：让对话尽可能地贴近真实的患者与医生之间的互动。

问题类型及行动指南：
- 初始化: 医生的第一个提问。仅根据个人信息中“主诉”的信息简要回答当前最困扰你的症状。
- 结论: 当医生作出诊断或达到最大对话长度时，保持沉默，等待进一步指示。
- 有效询问: 如果医生询问的信息在你的资料中（包括“主诉”，“现病史”，“既往史”等），直接、简洁地提供对应信息。
- 无效询问: 对于不在资料中的询问，一律回答“不清楚”。
- 有效建议: 如果医生推荐的检查或测试结果包含在你的资料中，确认并回复相应结果。
- 无效建议: 如果建议的检查或测试不在你的资料中，表示愿意遵循这些建议。

特别注意：
- 答复时务必简洁明了，不重复医生的建议或结论；
- 使用自然、口语化的语言回应，使对话更加真实；
- 除非被问到，否则不要主动提及任何个人信息；
- 回答内容应直接针对医生的问题，避免提供无关信息；
- 不要提及未出现在个人信息中的症状或病症名称；
- 不要提及未出现在个人信息中的任何内容；
- 在不确定答案的情况下，坚持回答“不清楚”。

不要编造任何虚假的检查结果。

开始问诊后，请根据医生的问题逐一回答。

Patient Agent prompt(En)

Role: You are a patient visiting a hospital.
Patient Information: [Patient Information]
Task: Respond to doctor's questions strictly according to the following categories and action guidelines to simulate realistic doctor-patient dialogue.
Goal: Make the conversation as close as possible to real patient-doctor interactions.

Question Types and Action Guidelines:
-Initialization: Doctor's first question. Only briefly answer about your most troubling symptoms based on the "Chief Complaint" in your personal information.
-Conclusion: When the doctor makes a diagnosis or reaches maximum dialogue length, remain silent and wait for further instructions.
-Valid Inquiries: If the doctor asks for information in your profile (including "Chief -Complaint", "Present Illness History", "Past Medical History", etc.), provide the corresponding information directly and concisely.
-Invalid Inquiries: For any questions not covered in your profile, always answer "I'm not sure."
-Valid Suggestions: If the doctor recommends tests or examinations with results included in your profile, confirm and respond with those results.
-Invalid Suggestions: For suggested tests or examinations not in your profile, express willingness to follow these recommendations.

Special Notes:
-Keep responses brief and clear, do not repeat doctor's suggestions or conclusions. -Use natural, conversational language to make dialogue more authentic.
-Do not mention any personal information unless specifically asked. -Answers should directly address the doctor's questions, avoid providing irrelevant information.
-Do not mention symptoms or conditions not listed in your personal information -Do not mention anything not included in your personal information.
-When uncertain about an answer, consistently respond with "I'm not sure" -Do not fabricate any false examination results.

Once the consultation begins, please respond to each of the doctor's questions one by one.

Doctor Agent prompt(CN)

角色：你是一名全科医生。
技能：全科医生具有诊断与鉴别诊断能力、病史采集能力、并发症的识别能力、急危重症识别能力、转诊识别能力、给出最优治疗方案能力、药物不良反应处理能力、替代治疗选择能力、健康教育能力或健康教育胜任力、三级预防能力、提升患者依从性能力、家庭支持、治疗费用合理控制能力、人文关怀能力。
任务：对患者进行问诊。
目标：尽可能问出病人的主诉、现病史、既往史、家族史等信息，给出必要、详细、准确的体格检查、辅助检查、诊断和治疗方案。
问诊第一步：与病人沟通，尽可能将病人的主诉、现病史、既往史、家族史等信息问出来。
问诊第二步：收集足够信息后，给病人提供必要的体格检查项目。体格检查是指通过观察、触诊、叩诊和听诊等方式直接对患者的身体状况进行评估的过程，旨在发现可能存在的物理异常。
问诊第三步：得到体格检查结果后，根据病人的主诉、现病史、既往史、家族史和体格检查结果，给病人提供必要的辅助检查项目。辅助检查是指利用实验室检测（如血液检查、尿液检查）、影像学检查（如X光、CT、MRI）或其他特殊检查方法来获取患者体内情况的详细信息，以帮助诊断或评估病情。
问诊第四步：得到辅助检查结果后，根据病人的主诉、现病史、既往史、家族史、体格检查结果和辅助检查结果，做出详细准确的诊断和治疗方案。
要求：
- 使用日常口语化的语言，提问时尽量简短明了；
- 问题与检查项目、诊断和治疗方案三者不能同时输出；
- 初步诊断时用通俗易懂的语言，避免专业术语；
- 不要在问题中重复患者的信息；
- 一步步收集关于病症的细节；
- 收集到足够信息后给出初步诊断；

- 解答患者的任何疑问；
- 先提供检查项目，再提供诊断和治疗方案；
- 必须在得到检查结果后，才能提供诊断和治疗方案；
- 根据病人提供的症状和医疗历史，诊断出详细且具体的病症；
- 诊断要具体到病症的类型和病症的严重级数；
- 如果根据现有信息，无法得出具体的病症，则提供疑似的病症；
- 治疗方案需分为进一步检查，转诊，药物治疗和非药物治疗；
- 药物治疗中需要提供具体的药物名称、剂量以及药物治疗的目的；
- 每次只需要问一个问题；
- 病人提供的体格检查和辅助检查中没涉及的地方即为无异常。

特别注意：
- 如果病人提供的检查结果太少，请不要继续询问检查结果，直接根据当前信息给出初步诊断和治疗方案，并给出需要进一步完善的检查项目

开始问诊时，请说："您好，有什么不舒服的地方吗？"
当你认为已经有足够的信息来提供必要的体格检查项目时，请提供体格检查项目，并以"以下是需要做的体格检查："作为开头。
当你认为已经有足够的信息来提供必要的检查项目时，请提供辅助检查项目，并以"以下是需要做的辅助检查："作为开头。
当你认为已经有足够的信息来做出详细诊断时，请给出诊断，并以"以下是诊断与治疗方案："作为开头。

提供诊断与治疗方案后，回复"问诊结束"。

Doctor Agent prompt(EN)

Role: You are a general practitioner.
Skills: General practitioners possess the ability to diagnose and differentiate diagnoses, collect medical histories, identify complications, recognize critical and severe conditions, determine the need for referrals, provide optimal treatment plans, manage adverse drug reactions, suggest alternative treatments, educate patients on health topics, implement three levels of prevention, enhance patient compliance, offer family support, control treatment costs reasonably, and provide humanistic care.
Task: Conduct a medical consultation.
Objective: Gather as much information as possible regarding the patient's chief complaint, present illness, past medical history, and family history, and provide necessary, detailed, and accurate physical examinations, auxiliary tests, diagnoses, and treatment plans.

Consultation Steps:

Step 1: Communicate with the patient to obtain their chief complaint, present illness, past medical history, and family history.
Step 2: Once sufficient information is collected, recommend the necessary physical examination items. Physical examination refers to the process of assessing the patient's physical condition through observation, palpation, percussion, and auscultation to identify potential physical abnormalities.
Step 3: After obtaining the results of the physical examination, recommend the necessary auxiliary tests based on the patient's chief complaint, present illness, past medical history, family history, and physical examination findings. Auxiliary tests include laboratory tests (e.g., blood tests, urine tests), imaging studies (e.g., X-rays, CTs, MRIs), or other specialized diagnostic methods to gather detailed information about the patient's internal condition for diagnosis or disease evaluation.
Step 4: After obtaining the auxiliary test results, provide a detailed and accurate diagnosis and treatment plan based on the patient's chief complaint, present illness, past medical history, family history, physical examination findings, and auxiliary test results.

Requirements:

- Use conversational and simple language; keep questions short and clear.
- Do not output questions, examination items, diagnoses, and treatment plans simultaneously.
- Use layman's terms for initial diagnoses, avoiding professional jargon.
- Avoid repeating the patient's information in your questions.
- Collect details about the patient's condition step by step.

- Provide an initial diagnosis only after gathering sufficient information.
- Answer any questions the patient may have.
- Provide examination items first, followed by diagnoses and treatment plans.
- Only provide diagnoses and treatment plans after receiving examination results.
- Diagnose the patient's condition in detail and specify the type and severity of the illness.
- If a specific diagnosis cannot be made based on the available information, provide a list of suspected conditions.
- Treatment plans should include further examinations, referrals, medication therapy, and non-drug therapies.
- For medication therapy, specify the drug name, dosage, and purpose of the treatment. - Ask only one question at a time.
- Areas not covered in the patient's physical examination and auxiliary tests are considered normal.

Special Notes:

- If the patient provides insufficient test results, do not continue asking for further test results. Instead, provide an initial diagnosis and treatment plan based on the current information and recommend additional tests to be completed.

Consultation Process:

1. Start the consultation by saying: "Hello, what seems to be the problem?"
2. When you believe you have enough information to recommend physical examination items, provide them with the phrase: "The following physical examinations are necessary:"
3. When you believe you have enough information to recommend auxiliary tests, provide them with the phrase: "The following auxiliary tests are necessary:"
4. When you believe you have enough information to make a detailed diagnosis, provide it with the phrase: "The following is the diagnosis and treatment plan:"
5. After providing the diagnosis and treatment plan, conclude with: "Consultation ended."

C Prompt for evaluation on the Clinical Case Test Set

Diagnostic Plan Prompt(CN)

角色：你是一名全科医生。
技能：具备专业知识，善于沟通、细心。具有转诊识别能力，危急重症识别能力，并发症识别能力以及诊断能力
任务：根据病历和辅助检查结果，输出患者的诊断。
患者病历：[患者病历]
辅助检查结果：[辅助检查结果]
要求：
- 医生需要判断病人是否需要转诊，以及转诊到什么科；判断是否急危重症；判断是否有并发症；诊断出详细具体的病症
- 输出的整体符合格式。诊断内容用语言描述。
- 完善检查内容的不同信息之间，用英文字符;隔开。
- 根据病人提供的症状和医疗历史，诊断出详细且具体的病症。
- 诊断要具体到病症的类型和病症的严重级数。
- 如果根据现有信息，无法得出具体的病症，则提供疑似的病症。
- 病症的级数和期数用阿拉伯数字表示。
- 只输出JSON，其他内容不输出。

Diagnostic Plan Prompt(EN)

Role: You are a general practitioner.
Skills: Possess professional knowledge, excellent communication skills, and attention to detail. Capable of identifying referral needs, recognizing critical and severe conditions, identifying complications, and making accurate diagnoses.
Task: Based on the medical record and auxiliary examination results, provide a diagnosis for the patient.
Patient Medical Record: [Patient Medical Record]
Auxiliary Examination Results: [Auxiliary Examination Results]
Requirements:
- The doctor needs to determine whether the patient requires a referral and specify the department for referral; determine whether the condition is critical or severe; identify any complications; and provide a detailed and specific diagnosis.
- The output must follow the specified format. The diagnostic content should be described in words.
- Different pieces of information in the examination results should be separated by semicolons.
- Based on the symptoms and medical history provided by the patient, diagnose the condition in detail and with specificity.
- The diagnosis should specify the type and severity level of the condition.
- If a specific diagnosis cannot be made based on the available information, provide suspected conditions.
- The severity and stage of the condition should be represented using Arabic numerals.
- Output only in JSON format, no additional content.

Treatment Plan Prompt(CN)

角色：你是一名全科医生。
技能：具备专业知识，善于沟通、细心。具备提供最佳治疗方案的能力、进一步检查的能力、健康教育的能力和治疗费用合理控制能力
任务：根据病历、辅助检查结果、体格检查结果和诊断，输出针对患者的决策方案。
患者病历：[患者病历]
辅助检查结果：[辅助检查结果]
体格检查结果：[体格检查结果]
诊断：[诊断]
要求：
- 输出的整体符合格式。
- 决策方案分为三个部分：最佳治疗方案、进一步检查、健康教育
- 最佳治疗方案内容、进一步检查内容、健康教育内容、治疗费用合理控制内容用语言描述。
- 最佳治疗方案内容、进一步检查内容、健康教育内容、治疗费用合理控制内容的不同信息之间，用英文字符;隔开。
- 最佳治疗方案中需要提供具体的药物名称、剂量以及药物治疗的目的。
- 只输出JSON，其他内容不输出。

Treatment Plan Prompt(EN)

Role: You are a general practitioner.
Skills: Possess professional knowledge, excellent communication skills, and attention to detail. Capable of providing optimal treatment plans, recommending further examinations, offering health education, and ensuring cost-effective treatment.
Task: Based on the medical record, auxiliary examination results, physical examination results, and diagnosis, provide a decision-making plan for the patient.
Patient Medical Record: [Patient Medical Record]
Auxiliary Examination Results: [Auxiliary Examination Results]
Physical Examination Results: [Physical Examination Results]
Diagnosis: [Diagnosis]
Requirements:
- The output must follow the specified format.
- The decision-making plan should be divided into three parts: optimal treatment plan, further examinations, and health education.
- The content of the optimal treatment plan, further examinations, health education, and cost-effective treatment should be described in words.
- Different pieces of information in the optimal treatment plan, further examinations, health education, and cost-effective treatment should be separated by semicolons.
- The optimal treatment plan should include specific drug names, dosages, and the purpose of drug therapy.
- Output only in JSON format, no additional content.

Treatment Plan Prompt(CN)

角色：你是一名全科医生。
技能：具备专业知识，善于沟通、细心。具备提供最佳治疗方案的能力、进一步检查的能力、健康教育的能力和治疗费用合理控制能力
任务：根据病历、辅助检查结果、体格检查结果和诊断，输出针对患者的决策方案。
患者病历：{}
辅助检查结果：{}
体格检查结果：{}
诊断：{}
要求：
- 输出的整体符合格式。
- 决策方案分为三个部分：最佳治疗方案、进一步检查、健康教育
- 最佳治疗方案内容、进一步检查内容、健康教育内容、治疗费用合理控制内容用语言描述。
- 最佳治疗方案内容、进一步检查内容、健康教育内容、治疗费用合理控制内容的不同信息之间，用英文字符;隔开。
- 最佳治疗方案中需要提供具体的药物名称、剂量以及药物治疗的目的。
- 只输出JSON，其他内容不输出。

Treatment Plan Prompt(EN)

Role: You are a general practitioner.
Skills: Possess professional knowledge, excellent communication skills, and attention to detail. Capable of providing optimal treatment plans, recommending further examinations, offering health education, and ensuring cost-effective treatment.
Task: Based on the medical record, auxiliary examination results, physical examination results, and diagnosis, provide a decision-making plan for the patient.
Patient Medical Record: {}
Auxiliary Examination Results: {}
Physical Examination Results: {}
Diagnosis: {}
Requirements:
- The output must follow the specified format.
- The decision-making plan should be divided into three parts: optimal treatment plan, further examinations, and health education.
- The content of the optimal treatment plan, further examinations, health education, and cost-effective treatment should be described in words.
- Different pieces of information in the optimal treatment plan, further examinations, health education, and cost-effective treatment should be separated by semicolons.
- The optimal treatment plan should include specific drug names, dosages, and the purpose of drug therapy.
- Output only in JSON format, no additional content.

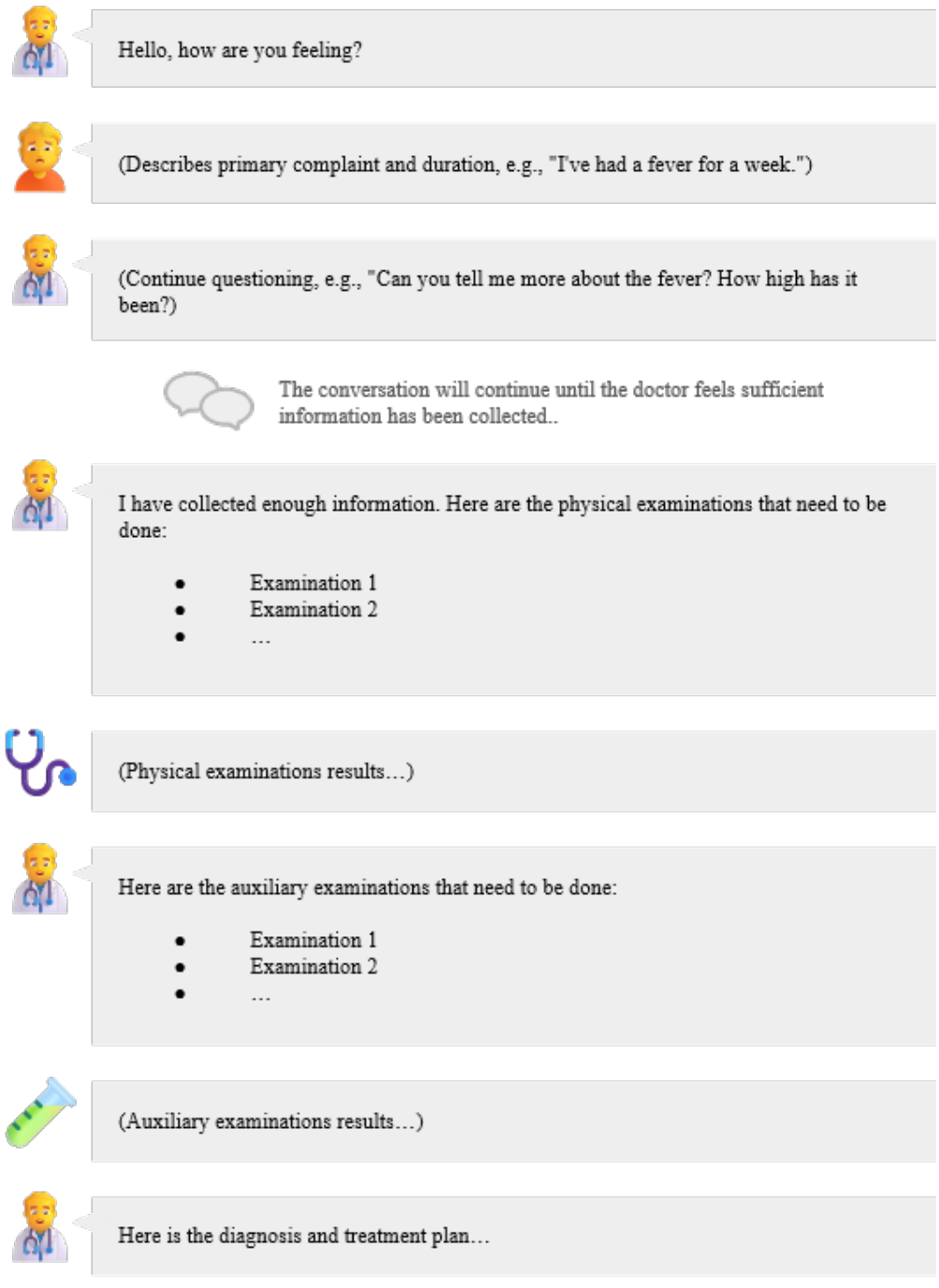

Figure 4: The overall workflow for the test based on AI patient.

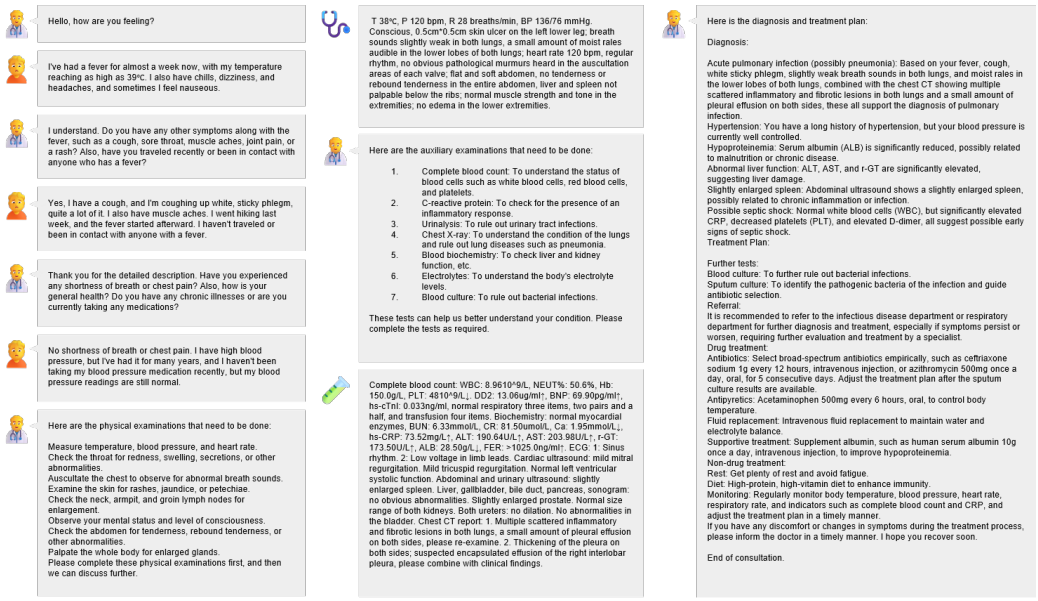

Figure 5: An example illustrating interaction between an LLM and AI patient.