

CLINICALMC: A Benchmark for Multi-Course Clinical Decision-Making with Large Language Models

Ruihui Hou[◇], Siyi Zhu[◇], Ziyue Huai[◇], Guangya Yu[◇], Yongqi Fan[◇],
Chunming Wang[♣], Tong Ruan^{◇*}

[◇]East China University of Science and Technology, Shanghai, China,
[♣]Renji Hospital Affiliated to Shanghai Jiaotong University
School of Medicine, Shanghai, China.

Abstract

Large language models (LLMs) have been widely adopted in healthcare, yet they still encounter significant challenges in complex clinical decision-making scenarios. Existing benchmarks primarily assess LLM performance in single-course settings and lack systematic evaluation in multi-course scenarios, where a patient’s condition evolves over time. To address this gap, we propose ClinicalMC, a benchmark for multi-course clinical decision-making. It includes 1,275 Chinese and 5,804 English samples across four stages from admission to discharge. These stages cover triage, first-course examination/diagnosis/treatment, subsequent multi-course examination/assessment/treatment, and final diagnosis. In ClinicalMC, patients in the English dataset undergo an average of 5.11 clinical courses, whereas those in the Chinese dataset undergo 3.42. To assess LLM performance, we construct a multi-agent evaluation framework that includes patient, examiner, and doctor agents. Based on the benchmark and framework, we design two experimental settings—a single-turn static setting and a multi-turn dynamic setting—and assess three categories of LLMs: 1) closed-source LLMs like GPT5-mini; 2) open-source LLMs like DeepSeek-V3.2, and 3) medical LLMs like HuatuoGPT-o1. Through extensive evaluation, we aim to better understand LLM performance in the medical domain and support its effective deployment in healthcare.¹

1 Introduction

Large language models (LLMs) have shown strong performance in various medical NLP tasks, including information extraction (Zhan et al., 2025), text generation (Lin et al., 2023) and question answering (Jin et al., 2021). However, their reliability remains limited in complex clinical decision-making

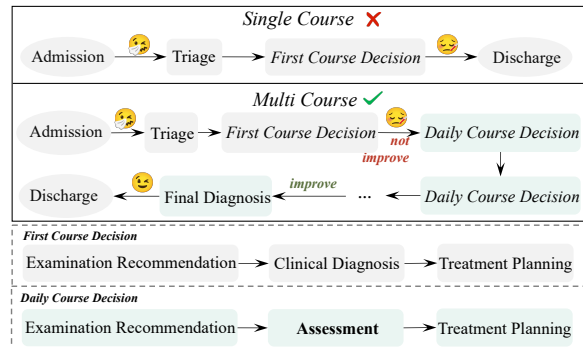


Figure 1: The solid boxes highlight the distinctions between our clinical decision-making tasks and previous benchmarks. Both “First Course Decision” and “Daily Course Decision” consist of three subtasks each, while the dashed boxes provide their detailed descriptions.

scenarios (Hager et al., 2024), which require the continuous integration of heterogeneous data (e.g., vital signs, laboratory results) and real-time reasoning under evolving patient conditions (Sutton et al., 2020). This limitation highlights the necessity of systematically evaluating the LLM applications in multi-course² clinical decision-making.

Clinical decision-making is a multi-stage, iterative process that often spans several treatment courses (Hager et al., 2024). Upon admission, clinicians first determine the most suitable department for each patient based on their primary presenting symptoms. During the first course, they gather relevant clinical information and recommend necessary examinations to guide preliminary diagnostic and treatment decisions. If the patient’s condition fails to improve, additional examinations are conducted in subsequent courses to reassess the clinical condition and promptly adjust the treatment plan. This iterative process continues until the patient’s condition stabilizes and discharge criteria are met. The

* Corresponding Authors.

¹Data and code are available at the URL <https://github.com/hzyuezh/ClinicalMCPD>.

²The course is a continuous record of a patient’s condition and treatment during hospitalization, including key details such as vital signs, surgeries, and major clinical changes.

Table 1: Overview of clinical decision-making benchmarks. “Dept.,” “CAS.,” “A2D.,” and “Multi-C.” stand for “department,” “assessment task”, “admission to discharge process”, and “multiple courses decision”, respectively. “continuous assessment” indicates whether the patient’s condition is continuously assessed.

Dataset	Data Source	# Dept.	A2D.	CAS.	Multi-C.	Language	# EHRs
MedR-Bench (Qiu et al., 2025)	PMC-patients	10	×	×	×	en	1,453
ClinicalLab (Yan et al., 2024)	Grade 3A Hospital	24	✓	×	×	zh	1,500
MIMIC-CDM (Hager et al., 2024)	MIMIC-IV	1	×	×	×	en	2,400
MedChain (Liu et al., 2025b)	Medical Website	19	✓	×	×	zh	12,163
AI Hospital (Fan et al., 2025b)	Medical Website	6	×	×	×	zh	506
CRAFT-MD (Johri et al., 2025)	MedQA	12	×	×	×	en	2,000
MAP (Chen et al., 2025)	MIMIC-IV	9	×	×	×	en	51,274
MedJourney (Wu et al., 2024)	Smart Hospital	12	×	×	×	zh	4,960
DiReCT (Wang et al., 2024)	MIMIC-IV	5	×	×	×	en	511
ClinicalMC (Ours)	MedEureka, PMC-patients	16, 24	✓	✓	✓	zh, en	1,275(zh), 5,804(en)

overall process is illustrated in Fig. 1.

Several benchmarks have been proposed for clinical decision-making, which can broadly be categorized into exam-based and clinical case-based benchmarks. Exam-based benchmarks, such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and MMLU (Hendrycks et al., 2021), primarily consist of Q&A pairs extracted from medical books and literature, aiming to evaluate the domain knowledge of LLMs. However, they are largely biased toward theoretical knowledge and fail to align with actual clinical decision scenarios. Clinical case-based benchmarks such as Clinicallab (Yan et al., 2024), AI Hospital (Fan et al., 2025b), and MedJourney (Wu et al., 2024) aim to simulate real-world clinical scenarios. However, they typically focus on single-course decision-making, involving only a single round of diagnosis and treatment, overlooking the crucial process of reassessing and adjusting treatment plans when a patient fails to improve across multiple courses. In this work, we further address this gap by modeling multi-course decision-making scenarios that better reflect real clinical practice. For ease of comparison, we summarize the differences between our benchmark and the most relevant clinical benchmarks in Table 1.

Hence, in the paper, we introduce **ClinicalMC**, a novel benchmark for evaluating the multi-course clinical decision-making capabilities of LLMs. To construct this benchmark, we collect clinical records that encompass multiple changes in patient conditions and incorporate condition assessment tasks into each key decision point throughout the clinical course. In addition, we design a three-round annotation workflow to ensure high-quality and consistent annotations. Using this approach,

we build 1,275 Chinese samples (covering 16 departments) and 5,804 English samples (covering 24 departments) from MedEureka (Fan et al., 2025a) and PMC-patients (Zhao et al., 2022). To facilitate systematic evaluation on ClinicalMC, we develop a multi-agent evaluation framework comprising a patient agent, an examiner agent, and a doctor agent. The patient agent provides the primary symptoms. The examiner agent provides feedback on the examination results. The doctor agent makes decisions at each stage of the workflow based on the patient’s evolving condition. Using this benchmark and framework, we construct two experimental settings—a single-turn static setting and a multi-turn dynamic setting—and conduct a comprehensive evaluation with a range of doctor agents, including closed-source LLMs such as GPT-4o-mini (Hurst et al., 2024), open-source LLMs such as DeepSeek-V3.2 (Liu et al., 2025a), and medical LLMs such as HuatuoGPT-o1 (Chen et al., 2023).

In summary, our contributions include:

- We introduce a novel benchmark for multi-course clinical decision-making, ClinicalMC. The benchmark comprises 1,275 Chinese samples across 16 departments and 5,804 English samples across 24 departments.
- The main characteristic of ClinicalMC is its inclusion of multiple clinical courses for each patient, enabling a more realistic representation of how a patient’s condition evolves over time. In the English dataset, patients have an average of 5.11 clinical courses, whereas in the Chinese dataset, the average is 3.42.
- We evaluate medical LLMs as well as closed- and open-source LLMs on ClinicalMC, in-

dicating that state-of-the-art medical models like instruction-tuned HuatuoGPT-o1(7B) achieve average performance of 43.40% and 47.77% on Chinese and English, respectively, far below human performance (85.00% and 87.51%). We further provide detailed analyses and suggest future research directions.

2 Problem Formulation

In this work, we evaluate the complete clinical process from patient admission to discharge. Each clinical task can be formally defined as:

Triage (TR): This task requires the doctor to select the most suitable department dp from a set of candidate departments ds , given the patient’s chief complaint cc and basic information bi . Formally, this is represented as: $dp = TR(cc, bi, ds)$.

Examination Recommendation (ER): This task involves predicting the necessary auxiliary examinations ex based on the patient’s chief complaint, present history ph_1 , past history ph_2 , and physical examination pe . Formally, this can be represented as: $ex = ER(cc, bi, ph_1, ph_2, pe, dp)$. For examination recommendations across multiple courses, the input includes the patient’s chief complaint of the current course, along with all prior patient information. This can be represented as: $ex' = ER(emr, pc, cc', pe')$, where cc' , ex' , and pe' represent the chief complaint, examination recommendation, and physical examination in the current course. pc and emr represent the previous course and the patient’s admission information.

Clinical Diagnosis (CD): This task requires the doctor to determine the patient’s preliminary diagnosis pd , the corresponding diagnostic basis pb , and the differential diagnosis dd , based on the patient’s chief complaint, present history, past history, physical examination, and auxiliary examinations. It can be formally represented as: $pd, pb, dd = CD(cc, bi, ph_1, ph_2, pe, dp, ex)$.

Assessment (AS): This task requires the doctor to assess the patient’s condition, based on the chief complaint and physical examination of the current course. The assessment may involve updating an existing diagnosis or introducing a new one. Formally, the task is defined as: $as' = AS(cc', pe', ex', emr)$, where as' represent the clinical assessment for the current course.

Treatment Planning (TP): This task involves predicting the optimal treatment plan based on the patient’s chief complaint, present history, past his-

tory, physical examination, auxiliary examinations, preliminary diagnosis, diagnostic basis, and differential diagnosis. It can be formally expressed as: $tp = TP(emr)$. For treatment planning across multiple courses, the input also includes the current course’s data. This can be represented as $tp' = TP(emr, cc', pe', ex', as')$.

Final Diagnosis (FD): This task requires the doctor to determine the final diagnosis fd and its supporting basis fb based on the entire clinical trajectory. This task can be formally represented as: $fd, fb = FD(emr, pn)$, where $pn = [pc_1, pc_2, \dots, pc_n]$ is the sequence of n courses. Each course $pc_i (1 \leq i \leq n)$ includes the chief complaint, physical examination, auxiliary examination, assessment, and treatment plan: $pc_i = (cc', ex', pe', as', tp')$.

3 ClinicalMC Construction

In this section, we provide a detailed description of the data collection and processing, quality control, and data statistics and analysis.

3.1 Data Collection and Processing

For the **Chinese data**, we use Electronic Health Records (EHRs) from MedEureka as the original data source. To obtain strictly anonymized and high-quality EHRs, we process the data in two stages. In the first stage, we identify EHRs containing personal information (e.g., names, phone numbers) using regular expressions, and replace sensitive data with placeholders (e.g., “Patient A”) or randomly generated values, resulting in 6,947 EHRs. In the second stage, we further filter the data to retain only complete and high-quality EHRs. We first remove EHRs lacking key information (e.g., chief complaints, diagnoses, or treatment processes), retaining 5,106 EHRs. We then exclude EHRs with a final outcome of death, leaving 4,179 EHRs, and finally eliminate duplicate records through fine-grained demographic matching (e.g., gender and occupation). After this stage, we obtain 3,317 high-quality EHRs, each containing multiple treatment courses. For the **English data**, we use 167,034 anonymized case reports from PMC-Patients as the original data source. To obtain high-quality multi-course reports, we conduct three screening steps. First, we use the GPT-4o model to remove reports that lack multiple courses or contain incomplete clinical courses (e.g., no improvement or death), retaining 37,357 reports. Sec-

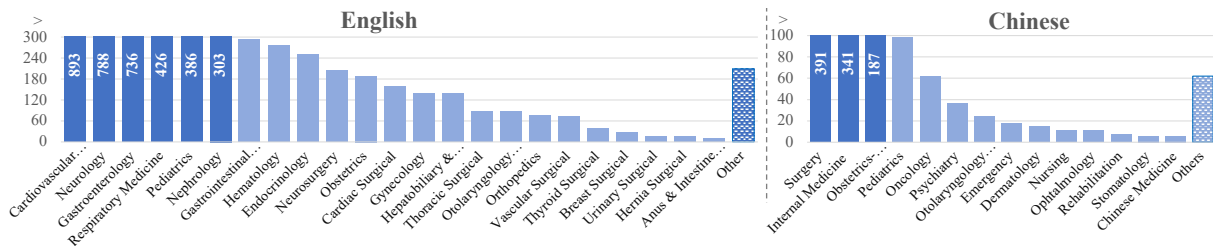


Figure 2: The department distribution of the Chinese and English datasets.

Table 2: Statistics of our constructed dataset.

Language	Avg. Notes	Max Notes	Min Notes	# EHRs
English	5.11	11	2	5,804
Chinese	3.42	10	2	1,275

ond, we remove reports missing key fields such as admission and final diagnosis, or those labeled as “undiagnosed”, leaving 15,572 reports. Finally, we exclude non-human data (e.g., treatment reports for animals). After this rigorous screening process, we ultimately retain 6,748 reports. Additionally, to ensure compliance with ethical standards, three clinicians from a Grade 3A hospital conduct a thorough ethical review of the final dataset, confirming that no ethical or moral guidelines are violated.

3.2 Quality Control

To construct ClinicalMC, we assemble a professional annotation team comprising three inspectors and two reviewers. The dataset is first automatically segmented from multi-course EHRs using an LLM. Subsequently, three clinically trained inspectors perform an initial verification, followed by a dual review conducted by two senior clinicians. The detailed annotation workflow is provided in Appendix A.1. After a rigorous two-stage quality review, we obtain 1,275 high-quality Chinese EHRs and 5,804 high-quality English EHRs. To further ensure data integrity and clinical relevance, we conduct an additional quality-control procedure involving three senior clinicians, each with over ten years of clinical experience and independent of the annotation reviewers. For this assessment, we randomly sample 3,000 cases from the English dataset (51.68%) and 1,000 cases from the Chinese dataset (78.43%). We design a standardized scoring framework that presents complete case information and requires clinicians to assess six binary quality dimensions: 1) rationality of course segmentation, 2) accuracy of triage, 3) correctness of diagnostic results, 4) appropriateness of treatment plans,

5) accuracy of clinical assessments, and 6) accuracy of examination recommendations. Clinicians make a “yes/no” judgment for each dimension, and a case is deemed valid only when all criteria are satisfied. Evaluation results show that 93.3% of sampled cases meet the predefined quality standards. The pass rates for individual criteria range from 91.9% to 96.3%, indicating consistently high overall quality. The Cohen’s kappa (Banerjee et al., 1999) for inter-reviewer agreement is 0.85, demonstrating strong consistency among reviewers. For the remaining 6.7% of cases that do not meet the standards, we perform manual corrections to ensure the reliability and completeness of the dataset.

3.3 Data Statistics and Analysis

We conduct an in-depth statistical analysis of clinical decision-making from two perspectives. 1) **Department distribution.** Fig. 2 presents the department distribution in both the Chinese and English EHR datasets. In the English dataset, the “Cardiovascular Medicine” department contains the most samples (893 EHRs), whereas the “Anus & Intestine Surgery” department has the fewest (10 EHRs). In the Chinese dataset, the “Surgery” department has the largest sample size (391 EHRs), while the “Chinese Medicine” and “Stomatology” department have the smallest, with only 6 EHRs each. By analyzing the department distribution in both datasets, we observe an imbalance in the sample sizes, reflecting the real-world situation in clinical data. This imbalance is likely due to differences in clinical demand across different departments. 2) **Number of courses.** As shown in Table 2, the English dataset has an average of 5.11 courses per patient, ranging from 2 to 11. In comparison, the Chinese dataset has an average of 3.42 courses per patient, with a range of 2 to 10.

4 Evaluation Framework

Inspired by AI Hospital (Fan et al., 2025b), we develop an evaluation framework, **SimHospital**,

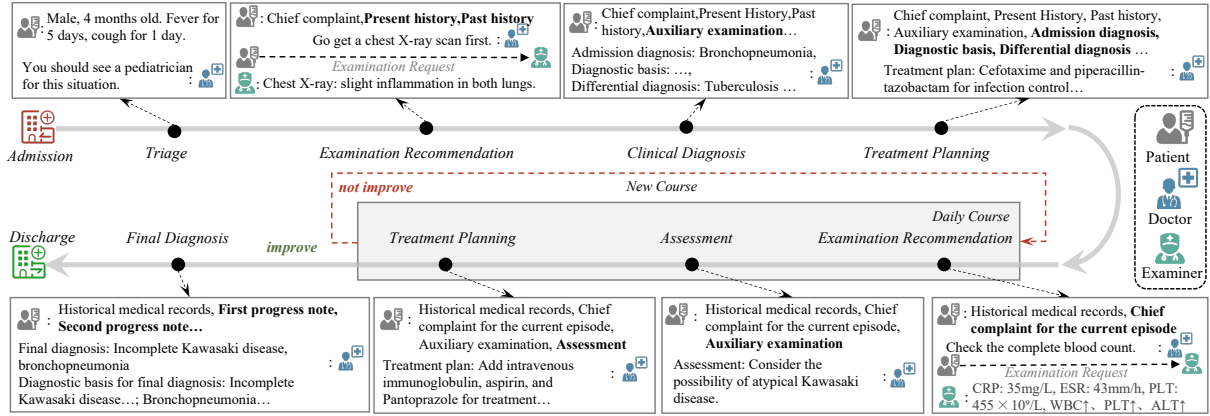


Figure 3: The SimHospital framework includes a doctor agent, an examiner agent, and a patient agent. In different tasks, different roles will engage in dialogues. When the patient shows improvement and is ready for discharge, the final diagnosis task is performed; otherwise, the patient continues into a new course. The **bold** text indicates the information that has been newly added in the current task compared to the previous task.

which consists of a doctor agent, a patient agent, and an examiner agent. GPT-4o-mini is used for the patient and examiner agents, while various LLMs are employed as the doctor agent to assess clinical decision-making performance. We also conduct an ablation study of the patient and examiner models in Appendix A.7.

4.1 Agent Behavior Setting for All Role

Examiner. The examiner agent is responsible for providing relevant examination results upon request from the patient agent. If the requested examination has corresponding results available, the examiner agent returns those results to the doctor. Otherwise, it responds with an indication that no such examination has been conducted. **Patient.** The patient agent’s main task is to interact with the doctor and the examiner agents. To match the actual situation, we add the chief complaint, present history, past history, and physical examination to the prompts of the patient agent, but do not specify the diagnosis or treatment plan. If the doctor suggests performing a specific examination, the patient agent follows the suggestion and provides the examination to the examiner agent. **Doctor.** The doctor agent’s primary task is to gather and analyze patient information to complete clinical decision-making tasks, including triage, examination recommendation, clinical diagnosis, assessment, treatment planning, and final diagnosis.

4.2 Clinical Workflow

The SimHospital Framework simulates the entire process from admission to recovery and discharge

by constructing multiple agents, as illustrated in Fig. 3. The interaction begins with the patient agent and proceeds through four stages. In the first stage, the patient agent presents a chief complaint, and the doctor agent recommends the appropriate department. In the second stage, the doctor interacts with both the patient and examiner agents, recommends necessary examinations, and makes clinical diagnoses and treatment plans. In the third stage, the patient agent enters the multi-course phase, during which the doctor agent sequentially performs tasks such as examination recommendation, assessment, and treatment planning based on the patient’s complaints for the current course. This process repeats until the patient recovers and is ready for discharge. In the fourth stage, the doctor agent provides a discharge diagnosis based on the patient’s complete medical information.

5 Experiments

In this section, we implement state-of-the-art models on our newly constructed ClinicalMC benchmark, aiming at assessing their performance and identifying the underlying challenges.

5.1 Experimental Setup

Baseline Model. We evaluate four categories of LLMs: 1) **Medical LLMs**, including MedGemma (Sellergren et al., 2025), Baichuan-M2 (Dou et al., 2025), HuatuoGPT-o1, and Apollo2-7B (Zheng et al., 2024). Additionally, as some medical LLMs demonstrate strong performance across different languages, we use HuatuoGPT2 (7B, 13B, and 34B) (Chen et al., 2023)

Table 3: Experimental results on English data (%). “T”, “E”, “PD”, “PB”, “DD”, “TP”, “FD”, and “FB” refer to triage, examination recall, preliminary diagnosis, preliminary diagnosis basis, differential diagnosis, treatment planning, final diagnosis, and final diagnosis basis, respectively. “CE”, “CA”, and “CT” represent the examination recommendation, assessment, and treatment planning for each course, respectively.

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>Medical LLMs</i>												
Apollo2-7B	61.06	72.76	29.93	61.99	40.95	5.37	37.13	23.98	10.06	65.27	74.61	43.92
Asclepius-Llama2-13B	0.02	0.00	0.00	44.05	38.30	1.42	0.00	22.66	1.02	0.00	31.25	12.61
Asclepius-Llama2-7B	0.02	0.00	0.00	43.65	38.45	1.42	0.00	22.49	1.01	0.00	31.29	12.58
MedGemma	63.02	19.88	22.40	76.99	62.79	10.07	24.66	50.80	2.10	70.38	85.85	44.45
Baichuan-M2	61.73	25.50	24.82	74.85	51.76	9.36	25.47	54.02	2.27	80.16	86.33	45.12
HuatuogPT-o1-7B	58.55	59.00	20.01	58.44	51.30	8.21	58.41	51.67	1.64	76.13	82.10	47.77
<i>Open-source LLMs</i>												
Llama-3.3-70B	63.66	17.33	20.51	71.97	51.83	10.87	15.60	70.16	5.74	82.19	79.81	44.52
Llama-3.2-3B	46.36	34.77	14.85	48.76	38.91	5.96	26.15	48.05	3.74	63.76	69.87	36.47
Mistral-7B-v0.3	37.35	59.37	16.47	65.53	46.47	8.45	45.89	58.97	4.85	72.96	79.37	45.06
Mixtral-8x22B	59.84	42.64	27.67	71.29	50.60	10.91	39.70	70.53	6.25	87.84	80.49	49.80
Falcon3-7B	50.76	43.81	14.67	60.57	47.12	7.47	42.08	65.48	4.36	67.57	76.51	43.67
Qwen2.5-72B	62.67	17.73	27.27	74.57	50.66	11.10	22.92	59.08	5.57	77.49	87.56	45.15
Qwen2.5-32B	62.79	19.06	22.13	74.42	52.38	11.72	24.72	66.11	6.39	79.75	86.26	45.98
Qwen2.5-14B	65.13	14.98	25.85	73.88	52.28	10.40	19.18	61.67	4.63	90.59	80.61	45.38
Qwen2.5-7B	59.30	52.15	10.94	59.79	49.03	8.13	39.32	55.46	5.36	77.36	77.98	44.98
Qwen3-Next-80B-A3B	57.15	22.96	29.67	77.91	57.23	12.40	17.16	59.05	4.12	79.56	84.35	45.59
DeepSeek-V3.2-Chat	62.90	20.68	27.69	81.82	60.60	10.95	23.49	42.87	5.28	69.62	86.28	44.74
DeepSeek-V3.2-Reason	53.11	13.95	28.08	78.43	56.17	12.89	13.04	57.68	5.88	70.85	87.17	43.39
<i>Closed-source LLMs</i>												
GPT-4o-mini	65.73	17.74	34.73	72.80	59.08	11.62	17.16	54.44	2.30	93.13	79.75	46.23
GPT-5-mini	61.71	28.16	31.86	83.33	61.71	13.20	25.84	60.01	4.71	58.46	86.65	46.88
Qwen-turbo	63.78	42.24	32.96	72.30	55.08	9.97	51.41	46.68	2.09	89.96	79.98	49.68
<i>Other Method</i>												
Human (sampling)	90.00	86.39	88.75	84.85	82.80	83.06	83.05	87.22	92.09	90.95	93.45	87.51

for Chinese datasets and Asclepius-Llama2 (7B and 13B) (Kweon et al., 2024a) for English datasets. Notably, Asclepius-Llama2 was trained on the PMC-Patients dataset, making it well-suited for assessing potential data leakage risks. 2) **Open-source LLMs**, including Falcon3-7B (Almazrouei et al., 2023), Qwen2.5 (ranging from 7B to 72B), DeepSeek-V3.2 (Chat and Reason) (Liu et al., 2025a), Llama-3.3-70B (Grattafiori et al., 2024), Llama-3.2-3B, Mistral-7B (Jiang et al., 2023), Mixtral-8x22B (Jiang et al., 2024), and Qwen3-Next-80B-A3B³. Since the Llama series models exhibit certain capabilities in processing Chinese, we also evaluate their performance on Chinese datasets. 3) **Closed-source LLMs**, such as GPT-4o-mini (Hurst et al., 2024), GPT-5-mini, and Qwen-turbo. 4) **Other Method**. We randomly select 100 samples and invite a medical student who does not participate in the data annotation process to answer the questions.

Evaluation Metrics. For the triage task, we use *Accuracy* (*Acc*) as the evaluation metric. For examination recommendation, we adopt *Recall*, and for disease diagnosis—covering both clinical and final diagnosis—we use the *F1* score. To eval-

uate examination and diagnosis entities, we construct a standardized synonym list by first collecting synonyms from the Medeureka_corpus (Fan et al., 2025a). To address terminological inconsistencies due to differences in model training data, we select the largest model from each of 10 LLM series to independently generate synonym lists. These are merged with the Medeureka corpus and refined by three clinicians to ensure consistency and clinical validity. For diagnosis basis, we employ an LLM to assess (1) whether the medical reasoning process is logically coherent, and (2) whether the provided evidence sufficiently and effectively supports the predicted diagnosis. The scores for preliminary and final diagnosis are denoted as *PB_Score* and *FB_Score*, respectively. The detailed evaluation prompts are provided in Appendix A.11. For differential diagnosis, we focus on evaluating whether the set of predicted differential diagnoses adequately covers other clinically significant potential conditions. This metric is denoted as *DD_Score*. The detailed evaluation prompts are provided in Appendix A.11. For the assessment and treatment planning in each clinical course, we follow the approach of MedChain (Liu et al., 2025b) by decomposing model outputs into structured clinical entities. We then compute the

³<https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct>

Table 4: Experimental results on Chinese data (%).

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>Medical LLMs</i>												
Apollo2-7B	52.71	29.05	33.85	66.65	35.01	5.40	29.39	46.54	0.98	25.82	54.75	34.56
HuatuoGPT2-7B	42.82	27.72	0.26	49.80	29.99	2.47	40.79	28.83	0.46	0.47	24.36	22.54
HuatuoGPT2-13B	45.13	26.63	0.00	48.26	29.87	1.99	52.94	23.99	0.06	0.00	37.69	24.23
HuatuoGPT2-34B	57.59	29.65	26.18	61.53	34.43	5.04	42.84	35.31	0.73	22.24	37.23	32.07
MedGemma	65.65	28.64	30.13	76.78	46.34	6.91	13.73	79.74	2.92	58.48	83.72	44.82
HuatuoGPT-o1-7B	65.73	30.93	33.31	70.48	39.36	5.78	17.24	71.77	2.89	62.55	77.38	43.40
Baichuan-M2	54.20	32.74	35.67	78.37	47.03	5.11	18.81	76.34	2.80	68.45	80.93	45.50
<i>Open-source LLMs</i>												
Llama-3.3-70B	52.39	35.52	33.72	73.04	41.11	6.38	25.28	50.81	1.97	29.12	67.94	37.93
Llama-3.2-3B	46.20	23.79	10.63	38.32	24.33	1.84	50.67	21.55	0.18	3.47	21.13	22.01
Mistral-7B-v0.3	39.14	23.90	14.52	34.15	22.78	2.28	24.60	23.67	0.49	13.84	27.75	20.65
Mixtral-8x22B	53.10	27.23	26.15	60.19	25.16	4.75	22.97	36.26	1.14	23.75	52.00	30.25
Falcon3-7B	41.10	29.05	12.60	34.78	22.71	1.47	37.09	23.62	0.41	6.93	24.85	21.33
Qwen2.5-72B-Chat	57.41	23.74	37.96	73.98	40.05	5.94	20.88	49.17	2.11	31.88	70.32	37.59
Qwen2.5-32B-Chat	59.92	23.06	34.98	76.61	43.45	7.67	18.24	60.09	2.84	30.69	64.42	38.36
Qwen2.5-14B-Chat	53.57	21.16	36.20	74.73	37.99	7.84	23.68	63.22	2.63	30.54	64.36	37.81
Qwen2.5-7B-Chat	62.35	28.87	35.35	72.49	38.53	6.34	31.33	50.66	1.40	28.90	64.75	38.27
Qwen3-Next-80B-A3B	68.31	28.53	31.25	82.46	61.30	6.38	11.98	84.89	2.75	63.80	88.17	48.17
DeepSeek-V3.2-Chat	67.06	25.46	30.87	77.04	63.44	7.57	15.55	77.94	3.07	54.61	79.22	45.62
DeepSeek-V3.2-Reason	59.50	23.94	37.98	77.10	58.65	6.86	13.67	81.02	3.24	67.36	81.08	46.40
<i>Closed-source LLMs</i>												
GPT-4o-mini	54.43	20.65	32.58	69.47	36.41	5.81	15.46	53.97	1.56	28.28	64.66	34.84
GPT-5-mini	59.62	25.92	12.00	83.76	49.48	6.76	9.66	84.34	4.19	44.60	82.16	42.04
Qwen-turbo	55.69	25.88	33.80	73.29	38.68	7.49	26.13	62.31	2.71	29.77	69.08	38.62
<i>Other Method</i>												
Human (sampling)	90.91	85.61	87.68	82.91	88.14	79.41	86.70	82.76	84.53	83.71	82.73	85.00

Intersection over Union (IoU) between these entities and the gold-standard key interventions. This metric emphasizes the coverage of critical clinical actions rather than surface-level wording, and is denoted as CA_IoU and CT_IoU , respectively.

Implementation Details. We design two experimental settings: a single-round static setting and a multi-round dynamic setting. In the former, the ground-truth annotations from preceding tasks are provided as inputs to subsequent tasks. In the latter, the model outputs of preceding tasks are used as inputs to subsequent tasks. Detailed implementation details are presented in Appendix A.2.

5.2 Main Results

We systematically evaluate all baseline LLMs on the ClinicalMC under the single-turn setting. The English and Chinese results are reported in Table 3 and Table 4, respectively. We select one representative model from each of the medical, closed-source, and open-source LLM categories for evaluation under the multi-turn experimental setting. The detailed results are presented in Appendix A.5.

From the Table 3 and Table 4, we notice that: 1) All LLMs perform poorly on both the English and Chinese datasets, leaving substantial room for improvement compared to human performance (Avg scores of 85.00% and 87.51%, respectively). The best-performing model achieves only 49.68% on the English dataset and 48.17% on the Chi-

nese dataset, highlighting the significant challenge posed by our ClinicalMC benchmark. 2) LLMs perform worse in multi-course settings compared to single-course settings. Specifically, Llama-3.3-70B achieves a TP_IoU of 6.38% on Chinese data and 10.87% on English data, outperforming the CT_IoU by 4.41% and 5.13%, respectively. This decline is primarily due to the increasing complexity of clinical information as the number of courses grows. Patients’ records often contain redundant or repeated examinations and treatments, making it more challenging for the model to accurately assess the current condition and generate up-to-date treatment plans in real-time. 3) Notably, although Asclepius-Llama2 is trained on the PMC-Patients dataset, it performs poorly on the English subset of ClinicalMC. Specifically, the 7B and 13B variants achieve Avg of only 12.58% and 12.61%, respectively. This is primarily because ClinicalMC reconstructs the medical records into reasoning tasks that require cross-trajectory information integration and explicit clinical decision-making, through multi-trajectory decomposition and multiple rounds of human review. In contrast, Asclepius-Llama2 focuses more on medical record generation and local semantic modeling, limiting its effectiveness in such complex clinical reasoning scenarios. Consequently, these results further highlight the challenging nature of ClinicalMC for evaluating clinical

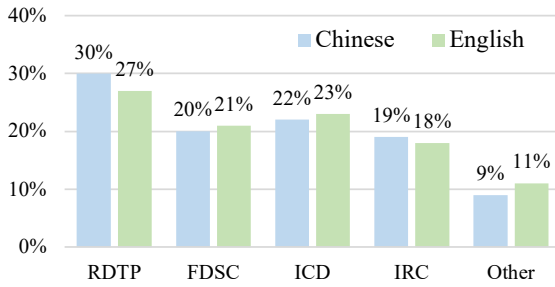


Figure 4: Distribution of error types.

reasoning capabilities.

5.3 Error Type

To guide future research in clinical decision-making for LLMs, we manually analyze and classify 200 error samples generated by LLMs on the Chinese and English datasets of ClinicalMC. These errors are categorized into five types: (a) **Redundant Diagnostic and Treatment Plan (RDTP)**: The model generates an excessive number of unnecessary diagnostic tests and treatment plans. (b) **Failure to Detect Subtle but Critical Changes (FDSC)**: The model fails to recognize subtle yet clinically significant changes in a patient’s condition—such as slight fluctuations in laboratory results—which may lead to delayed or inappropriate adjustments in diagnosis or treatment plans. (c) **Incorrect Clinical Diagnosis (ICD)**: Due to a lack of domain-specific medical knowledge or misinterpretation of clinical information, the model produces incorrect diagnostic conclusions. (d) **Incorrect Reasoning Chain (IRC)**: The diagnostic rationale produced by the model does not align with the actual clinical condition. (e) **Other Errors**: all other cases of errors. The error distribution is shown in Fig. 4, with illustrative Chinese and English examples included in Appendix A.8.

6 Related Work

Clinical decision-making benchmark. Clinical decision-making tasks refer to assisting doctors in making the most appropriate diagnosis and treatment decisions by continuously analyzing the patient’s chief complaints, medical history, examination results and other information (Hager et al., 2024). Existing clinical decision-making benchmarks can be broadly classified into two types: exam-based and clinical case-based benchmarks. Exam-based benchmarks include datasets such as MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), and

MMLU (Hendrycks et al., 2021), primarily consist of Q&A pairs extracted from medical books and literature. However, there is still a certain gap between these benchmarks and actual clinical decision-making. Therefore, recent studies have proposed benchmarks based on clinical case benchmarks, such as MedChain (Liu et al., 2025b), Clinicallab (Yan et al., 2024), MSDiagnosis (Hou et al., 2024), Ai-Hospital (Fan et al., 2025b), and MedJourney (Wu et al., 2024). However, these benchmarks mostly focus on a single course or simulate decision-making in outpatient settings. For example, Clinicallab evaluates tasks such as department guidance, clinical diagnosis, and treatment planning, but does not involve continuous assessment of patients after treatment until they recover and are discharged. Therefore, this study focuses on the clinical decision-making performance of models in multi-course after patient admission.

Agent for medical decision-making. Research on intelligent agents for medical decision-making can be divided into single-agent (Li et al., 2024a; Chen et al., 2024; Hou et al., 2026) and multi-agent (Tang et al., 2024; Li et al., 2024b) methods. In single-agent research, CoD (Chen et al., 2024) assesses potential candidate diseases by planning to inquire about the patient’s latent symptoms and generates a diagnostic chain from symptoms to possible diseases. In multi-agent research, medical decision-making problems are typically tackled through a multi-agent task division and collaboration paradigm, such as in frameworks like MDAgents (Kim et al., 2024), MedAgents (Tang et al., 2024), and Agent Hospital (Li et al., 2024b). MDAgents is a multi-agent framework that utilizes adaptive decision-making mechanisms to tackle medical decision-making challenges. It operates through multiple phases, including analyzing problem complexity, dynamically recruiting experts, and employing reasoning and decision-making processes at various stages to solve medical Q&A.

7 Conclusion

We introduce ClinicalMC, a benchmark comprising both Chinese and English datasets that encompass the full patient journey from admission to discharge. These stages cover triage, first-course examination/diagnosis/treatment, subsequent multi-course examination/assessment/treatment, and final diagnosis. To evaluate model performance in multi-course clinical decision-making, we develop a multi-agent

framework involving patient, examiner, and doctor agents. Based on the dataset and framework, we define two experimental settings—single-turn and multi-turn—and evaluate medical LLMs as well as closed-source and open-source LLMs and conduct extensive experimental analysis. The results show that ClinicalMC is a challenging dataset that warrants further research and exploration.

Limitations

This paper has two primary limitations that offer avenues for future research: First, the lack of multimodal information. The raw data used in this study primarily consist of textual medical records collected during hospitalization and do not cover multimodal data across multiple courses, such as medical images and time-series physiological signals. In future work, we plan to investigate the integration of medical imaging with text-based reasoning to support clinical decision-making over heterogeneous, multi-source data. Second, an imbalanced department distribution. The current dataset is mainly derived from a single data source, leading to imbalanced distributions across clinical departments. Although this imbalance partially reflects real-world clinical practice, it may still affect the model’s generalization performance in underrepresented departments. In future work, we will incorporate data from multiple healthcare systems to expand coverage and mitigate department-level imbalance.

Ethical Consideration

Our ClinicalMC benchmark is based on PMC-Patients and MedEureka, licensed under the Creative Commons Attribution 4.0 License. Accordingly, we assign the copyright of ClinicalMC to the CC-BY 4.0 license. In addition, we have meticulously reviewed our dataset to ensure it does not contain any harmful content, including gender bias, racial discrimination, or inappropriate material.

Acknowledgments

This paper was supported by the Shanghai Science and Technology Innovation Action Plan in Computational Biology (No. 24JS2840200).

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru,

Mérouane Debbah, Étienne Goffinet, Daniel Hessel, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23.

Junying Chen, Chi Gui, Anningzhe Gao, Ke Ji, Xidong Wang, Xiang Wan, and Benyou Wang. 2024. Cod, towards an interpretable medical agent using chain of diagnosis. *arXiv preprint arXiv:2407.13301*.

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

Zhen Chen, Zhihao Peng, Xusheng Liang, Cheng Wang, Peigan Liang, Linsheng Zeng, Minjie Ju, and Yixuan Yuan. 2025. Map: Evaluation and multi-agent enhancement of large language models for inpatient pathways. *arXiv preprint arXiv:2503.13205*.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, et al. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Yongqi Fan, Nan Wang, Kui Xue, Jingping Liu, and Tong Ruan. 2025a. [MedEureka: A medical domain benchmark for multi-granularity and multi-data-type embedding-based retrieval](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2825–2851, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025b. [AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

YanJun Gao, Timothy Miller, Dongfang Xu, Dmitriy Dligach, Matthew M Churpek, and Majid Afshar. 2022. Summarizing patients’ problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2022, page 2979.

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

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob

- Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Ruihui Hou, Shencheng Chen, Yongqi Fan, Lifeng Zhu, Jing Sun, Jingping Liu, and Tong Ruan. 2024. Msdiagnosis: An emr-based dataset for clinical multi-step diagnosis. *arXiv preprint arXiv:2408.10039*.
- Ruihui Hou, Dongge Xue, Hongli Sun, Ping He, Weiyan Zhang, and Tong Ruan. 2026. Cdaflow: Enhancing llm clinical decision-making through agentic workflow. *Expert Systems with Applications*, page 131806.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. *Mistral 7b*. Preprint, arXiv:2310.06825.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Leandra A Barnes, Hong-Yu Zhou, Zhuo Ran Cai, Eliezer M Van Allen, David Kim, et al. 2025. An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine*, pages 1–10.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. *MDAgents: An adaptive collaboration of LLMs for medical decision-making*. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Sunjun Kweon, Junu Kim, Jiyoung Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024a. *Publicly shareable clinical large language model built on synthetic clinical notes*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5148–5168, Bangkok, Thailand. Association for Computational Linguistics.
- Sunjun Kweon, Junu Kim, Jiyoung Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024b. *Publicly shareable clinical large language model built on synthetic clinical notes*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5148–5168, Bangkok, Thailand. Association for Computational Linguistics.
- Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. *MMedAgent: Learning to use medical tools with multi-modal agent*. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, Miami, Florida, USA. Association for Computational Linguistics.
- Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. 2024b. *Agent hospital: A simulacrum of hospital with evolvable medical agents*. *arXiv preprint arXiv:2405.02957*.
- Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2023. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.
- Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. 2025a. *Deepseek-v3. 2: Pushing the frontier of open large language models*. *arXiv preprint arXiv:2512.02556*.
- Jie Liu, Wenxuan Wang, Zizhan Ma, Guolin Huang, SU Yihang, Kao-Jung Chang, Haoliang Li, Linlin Shen, Michael Lyu, and Wenting Chen. 2025b. *Med-chain: Bridging the gap between LLM agents and*

- clinical practice with interactive sequence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Weiwei Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on real-world clinical cases. *arXiv preprint arXiv:2503.04691*.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cian Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17.
- Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 599–621, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic identification. *Journal of the American Medical Informatics Association*, 14(5):550–563.
- Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. DireCT: Diagnostic reasoning for clinical notes via large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin Wang, Jie Yang, et al. 2024. Med-journey: Benchmark and evaluation of large language models over patient clinical journey. *Advances in Neural Information Processing Systems*, 37:87621–87646.
- Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, et al. 2024. Clin-callab: Aligning agents for multi-departmental clinical diagnostics in the real world. *arXiv preprint arXiv:2406.13890*.
- Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. 2025. Ramie: retrieval-augmented multi-task information extraction with large language models on dietary supplements. *Journal of the American Medical Informatics Association*, 32(3):545–554.
- Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.
- Guorui Zheng, Xidong Wang, Juhao Liang, Nuo Chen, Yuping Zheng, and Benyou Wang. 2024. Efficiently democratizing medical llms for 50 languages via a mixture of language family experts. *arXiv preprint arXiv:2410.10626*.

A Appendix

A.1 Data Annotation

To create a high-quality benchmark, we organize a professional team of three inspectors and two reviewers, all trained in specialized medical knowledge. The annotation procedure includes first-round annotation, second-round checking, and third-round review.

First-round annotation. Since PMC-Patients contains case reports instead of complete EHRs, we use the GPT-4o model to convert these summaries into full EHRs containing multiple courses. For both **Chinese** and **English** datasets, the EHRs contain key information such as the primary and final diagnoses, but still lack the initial treatment plan and multiple progress notes. Therefore, we first input the patient’s primary diagnosis, auxiliary examinations, and chief complaint into the GPT-4o model to generate the initial treatment plan. We then prompt the model to segment the treatment process into multiple progress notes based on **temporal information** and **clinical status changes**, with each note following the standard SOAP format (Gao et al., 2022; Wang et al., 2024). To mitigate hallucinations during generation, we explicitly instruct the model in the prompt to avoid generating clinical information (such as examination results and time intervals) not present in the record summary. Additionally, inspired by Asclepius (Kweon et al., 2024b), we evaluate the similarity between the converted English and real EHRs using perplexity. Specifically, we fine-tune LLaMA-7B (Touvron et al., 2023) on 57,000 real discharge summaries from the MIMIC-III database (Johnson et al., 2016). We then measure the perplexity of 500 discharge summaries from two hospital datasets—MIMIC-IV (Johnson et al., 2023) and i2b2 (Uzuner et al., 2007)—as well as 500 case reports from PMC-Patients using the same model. Finally, we evaluate the perplexity of clinical cases synthesized from the PMC-Patients. Results show perplexity scores of 3.144 for MIMIC-IV and 5.916 for i2b2. In comparison, the original PMC-Patients data yields 72.471, while the GPT-4o-converted EHRs achieve a much lower score of 6.064. These results indicate that our synthetic notes are substantially more coherent and closely aligned with real hospital data.

Second-round checking. We invite a review team of three clinically trained medical students to perform quality checks on the annotations gener-

ated by GPT-4o. Any sample unanimously deemed invalid by all three reviewers is directly discarded. If only one or two reviewers raise concerns, the sample is manually re-annotated and retained only after all three reviewers agree that the revised annotation is appropriate. The criteria for determining annotation validity focus on three key aspects: 1) Consistency in the number of course records compared with the original EHR, including whether the model hallucinates nonexistent entries or omits essential course records; 2) Completeness and accuracy of examination information, ensuring that key results (e.g., laboratory findings) appear in the generated course records and remain faithful to the original data; 3) Correctness of field semantics, such as ensuring that the “chief complaint” reflects the patient’s subjective description of symptoms rather than objective examination findings. Additionally, we employ a batch-based iterative validation mechanism: each batch must achieve over 90% accuracy in the aggregated evaluation of the three reviewers before progressing to the next stage. This process effectively filters out structural inconsistencies, hallucinated content, and medical reasoning errors in the synthetic data, thereby establishing a reliable foundation for subsequent expert review.

Third-round review. We submit the preliminarily inspected EHRs to two clinicians for dual expert review. The clinicians randomly sample 30% of the cases for quality assessment and systematically evaluate whether each case narrative aligns with real clinical workflows (e.g., examination sequences, diagnostic reasoning, and treatment decisions) and whether any medical inaccuracies or potential safety risks are present. Any sample deemed unsatisfactory is returned to the previous stage for revision by the inspection team and then resubmitted for expert review. We repeatedly implement this iterative cycle of *expert feedback* → *manual correction* → *re-review* until the sampling accuracy consistently reaches 95% or higher. After multiple rounds of iteration and dual clinical review, we ultimately obtain 1,275 high-quality Chinese cases and 5,804 high-quality English cases. All cases pass rigorous evaluations of medical consistency, data safety, and factual accuracy.

Contributors. Medical students and clinicians are primarily recruited from the internship programs and clinical departments of a Grade 3A hospital and jointly participate in the data-annotation process. Compensation is provided at \$5–20/hr for medical students and \$50–100/hr for clinicians,

Table 5: The evaluation of LLM performance on ClinicalMC English data using GPT-4, with a maximum score of 10 points. ‘‘Comp.’’, ‘‘Prof.’’ and ‘‘Auth.’’ denote ‘‘Comprehensiveness’’, ‘‘Professionalism’’ and ‘‘Authenticity’’, respectively.

Model	Comp.	Prof.	Auth.	Safety	Total
<i>Medical LLMs</i>					
Apollo2-7B	1.81	3.05	2.40	1.00	8.26
Asclepius-Llama2-13B	1.18	2.36	1.66	1.00	6.20
Asclepius-Llama2-7B	1.13	2.11	1.49	1.00	5.73
MedGemma	1.06	2.06	1.16	1.00	5.28
Baichuan-M2	1.09	2.17	1.27	1.00	5.53
HuatuoGPT-o1-7B	1.01	1.97	1.08	1.00	5.06
<i>Open-source LLMs</i>					
Llama-3.3-70B	1.94	3.43	2.70	1.00	9.07
Llama-3.2-3B	1.55	2.70	2.14	1.00	7.39
Mistral-7B-v0.3	1.96	3.33	2.59	1.00	8.88
Mixtral-8x22B	1.96	3.35	2.72	1.00	9.03
Falcon3-7B	1.92	3.21	2.57	1.00	8.70
Qwen2.5-72B	1.95	3.47	2.84	1.00	9.26
Qwen2.5-32B	1.93	3.36	2.78	1.00	9.12
Qwen2.5-14B	1.97	3.52	2.80	1.00	9.29
Qwen2.5-7B	1.93	3.45	2.74	1.00	9.12
Qwen3-Next-80B-A3B	1.23	2.63	1.85	1.00	6.71
DeepSeek-V3.2-Chat	1.96	3.59	2.81	1.00	9.36
DeepSeek-V3.2-Reason	1.07	2.16	1.38	1.00	5.61
<i>Closed-source LLMs</i>					
GPT-4o-mini	1.98	3.44	2.78	1.00	9.20
GPT-5-mini	1.42	3.08	2.10	1.00	7.60
Qwen-turbo	1.97	3.41	2.68	1.00	9.06

based on task difficulty and required expertise.

A.2 Implementation Details

In this paper, we adopt two experimental settings, with all experiments conducted under a zero-shot setting. In the first experimental setting, for downstream tasks in the workflow, we provide the ground-truth annotations from preceding tasks as inputs, rather than using the model-generated outputs. In the second experimental setting, model responses from earlier tasks are directly used as inputs for subsequent tasks. To enhance the stability and reliability of the results and reduce the impact of randomness, each experiment is repeated three times, and the average performance is reported. For all experiments, the model temperature is set to 0.01. All experiments are conducted on four NVIDIA A800 GPUs (80 GB). For the open-source and medical LLMs, we deploy them using the vLLM framework⁴. For closed-source LLMs and DeepSeek-V3 and DeepSeek-R1, we use their official APIs⁵ for evaluation due to their excessively large parameter sizes.

⁴<https://github.com/vllm-project/vllm>

⁵<https://platform.DeepSeek.com/usage>

Table 6: The evaluation of LLM performance on ClinicalMC Chinese data using GPT-4.

Model	Comp.	Prof.	Auth.	Safety	Total
<i>Medical LLMs</i>					
Apollo2-7B	1.30	2.83	2.20	0.94	7.27
HuatuoGPT2-7B	1.30	2.94	2.24	0.92	7.43
HuatuoGPT2-13B	1.30	2.78	2.17	0.92	7.17
HuatuoGPT2-34B	1.54	2.93	2.36	1.00	7.83
MedGemma	1.79	3.37	2.76	1.00	8.92
HuatuoGPT-o1-7B	1.58	3.08	2.54	1.00	8.20
Baichuan-M2	1.73	3.30	2.68	1.00	8.71
<i>Open-source LLMs</i>					
Llama-3.3-70B	1.69	3.16	2.65	1.00	8.50
Llama-3.2-3B	1.21	2.33	1.86	0.97	6.37
Mistral-7B-v0.3	1.08	1.85	1.39	0.77	5.09
Mixtral-8x22B	1.48	2.78	2.32	1.00	7.58
Falcon3-7B	1.16	2.07	1.63	0.85	5.71
Qwen2.5-72B	1.80	3.41	2.78	1.00	8.99
Qwen2.5-32B	1.72	3.36	2.70	1.00	8.78
Qwen2.5-14B	1.69	3.31	2.67	1.00	8.67
Qwen2.5-7B	1.58	3.14	2.55	1.00	8.27
Qwen3-Next-80B-A3B	1.89	3.55	2.87	1.00	9.31
DeepSeek-V3.2-Chat	1.92	3.64	2.90	1.00	9.46
DeepSeek-V3.2-Reason	1.87	3.51	2.85	1.00	9.23
<i>Closed-source LLMs</i>					
GPT-4o-mini	1.66	3.24	2.63	1.00	8.53
GPT-5-mini	1.88	3.84	2.67	1.00	9.39
Qwen-turbo	1.71	3.37	2.69	1.00	8.77

A.3 LLM Evaluation

In this section, we primarily use GPT-4 to evaluate the performance of LLMs on ClinicalMC. The evaluation includes tasks such as preliminary diagnosis basis, differential diagnosis, first treatment plan, assessment and treatment in the multi-course, and final diagnosis basis. To account for potential instability in GPT-4’s responses, we conduct three evaluations for each model on each benchmark and calculate the average score. The specific prompts used are shown in Fig. 5. The experimental results of Chinese data and English data are shown in Table 6 and Table 5, respectively. The experimental results show that DeepSeek-V3 performs the best on both Chinese and English data. Specifically, DeepSeek-V3 achieves a *Total* score of 9.36 on English data and 9.46 on Chinese data.

A.4 Human Evaluation

In this section, to evaluate the quality and accuracy of the model’s decision results, we invite three medical experts with over ten years of clinical experience for manual evaluation. We randomly select 50 Chinese and 50 English EHRs, with each EHR anonymized to ensure that the evaluators cannot identify the model used. Furthermore, each EHR is evaluated by two different experts in a double-blind

Table 7: Human evaluation of LLM performance on ClinicalMC English dataset, with a maximum score of 10 points.

Model	Comp.	Prof.	Auth.	Safety	Total
<i>Medical LLMs</i>					
Apollo2-7B	1.40	2.20	2.00	1.00	6.60
Asclepius-Llama2-13B	1.10	2.10	1.80	1.00	6.00
Asclepius-Llama2-7B	1.00	2.00	1.60	1.00	5.60
MedGemma	1.30	3.30	2.20	1.00	7.80
Baichuan-M2	1.20	3.20	2.20	1.00	7.60
HuatuoGPT-o1-7B	1.20	3.10	2.00	1.00	7.30
<i>Open-source LLMs</i>					
Llama-3.3-70B	1.90	3.10	2.30	1.00	8.30
Llama-3.2-3B	1.30	2.30	2.10	1.00	6.70
Mistral-7B-v0.3	1.90	3.00	2.10	1.00	8.00
Mixtral-8x22B	1.90	3.10	2.20	1.00	8.20
Falcon3-7B	1.80	3.00	2.10	1.00	7.90
Qwen2.5-72B	1.90	3.30	2.60	1.00	8.80
Qwen2.5-32B	1.90	3.10	2.40	1.00	8.40
Qwen2.5-14B	1.80	3.30	2.50	1.00	8.60
Qwen2.5-7B	1.80	3.20	2.40	1.00	8.40
Qwen3-Next-80B-A3B	1.60	3.40	2.60	1.00	8.60
DeepSeek-V3.2-Chat	1.90	3.50	2.60	1.00	9.00
DeepSeek-V3.2-Reason	1.30	3.10	2.20	1.00	7.60
<i>Closed-source LLMs</i>					
GPT-4o-mini	1.90	3.20	2.40	1.00	8.60
GPT-5-mini	1.60	3.30	2.50	1.00	8.20
Qwen-turbo	1.90	3.20	2.30	1.00	8.50

cross-assessment setup. The evaluators score the decision results based on four dimensions: comprehensiveness, professionalism, authenticity, and safety. The scoring criteria align with the LLM evaluation standards outlined in Section LLM Evaluation. The manual evaluation results on the Chinese and English datasets are shown in Table 8 and Table 7, respectively. The table shows that the DeepSeek-V3 model performs the best, which is similar to the ranking obtained from the LLM evaluation in Appendix A.3. Specifically, on the English data, the DeepSeek-V3 model achieves a *Total* score of 9.00, while on the Chinese data, the *Total* score reaches 8.75.

A.5 Evaluation in Multi-turn Dynamic Environment

To evaluate model performance in dynamic environments, we use the responses generated by the models in previous tasks as input for subsequent stages of the clinical workflow, rather than relying on ground-truth answers. Specifically, based on the results of the static evaluation, we select representative models for comparison: Huatuo2-34B, Qwen2.5-7B, and GPT-4o-mini for the Chinese dataset; and Apollo2-7B, Mixtral-8X22B, and Qwen-turbo for the English dataset. Furthermore,

Table 8: The human evaluation of LLM performance on ClinicalMC Chinese data.

Model	Comp.	Prof.	Auth.	Safety	Total
<i>Medical LLMs</i>					
Apollo2-7B	1.00	2.70	2.00	1.00	6.70
HuatuoGPT2-7B	1.10	2.75	1.85	1.00	6.70
HuatuoGPT2-13B	1.00	2.55	1.70	1.00	6.25
HuatuoGPT2-34B	1.05	2.65	1.95	1.00	6.65
MedGemma	1.20	2.75	2.05	1.00	7.00
HuatuoGPT-o1-7B	1.00	2.30	1.95	1.00	6.25
Baichuan-M2	1.40	2.95	2.20	1.00	7.55
<i>Open-source LLMs</i>					
Llama-3.3-70B	1.15	2.90	2.00	1.00	7.05
Llama-3.2-3B	1.75	3.00	2.10	1.00	7.85
Mistral-7B-v0.3	1.00	2.20	1.70	1.00	5.90
Mixtral-8x22B	1.05	2.60	1.95	1.00	6.60
Falcon3-7B	1.00	2.35	1.65	1.00	6.00
Qwen2.5-72B	1.80	3.55	2.25	1.00	8.60
Qwen2.5-32B	1.50	3.00	2.05	1.00	7.55
Qwen2.5-14B	1.30	3.00	2.00	1.00	7.30
Qwen2.5-7B	1.15	3.00	2.00	1.00	7.15
Qwen3-Next-80B-A3B	1.55	3.00	2.25	0.90	7.70
DeepSeek-V3.2-Chat	1.95	3.25	2.55	1.00	8.75
DeepSeek-V3.2-Reason	1.25	2.80	2.30	1.00	7.35
<i>Closed-source LLMs</i>					
GPT-4o-mini	1.30	3.00	2.05	1.00	7.35
GPT-5-mini	1.55	3.25	2.30	1.00	8.10
Qwen-turbo	1.95	3.00	2.55	1.00	8.50

due to the inherent uncertainty in disease progression within dynamic settings—resulting in unpredictable task sequence lengths—we randomly sample 100 cases for experimentation. During evaluation, each stage’s output is compared against the gold standard. If an error occurs in any previous task, all subsequent tasks for that case are marked as invalid, simulating the cascading effect of errors in real-world applications. The experimental results are summarized in Table 9. From the table, we observe that all models experience a decline in performance in dynamic settings on the English dataset, primarily because early-stage errors often propagate downstream, negatively affecting later decisions. Specifically, on the English dataset, Apollo2-7B, Mixtral-8x22B, and Qwen-turbo show performance drops of 13.77%, 13%, and 9.06%, respectively, compared to static evaluation. However, we observe an opposite trend on the Chinese dataset, where performance slightly improves under dynamic evaluation. Specifically, HuatuoGPT2-34B, Qwen2.5-7B, and GPT-4o-mini achieve gains of 4.26%, 1.07%, and 3.55%, respectively. This discrepancy can be attributed to differences in data distribution and task complexity between the Chinese and English settings. Specifically, Chinese cases tend to involve shorter clinical

Table 9: Results under the Multi-turn Experimental Setting on Chinese and English Data (%).

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>English</i>												
Apollo2-7B	55.00	47.50	27.06	53.00	31.40	4.74	33.35	23.70	1.62	8.45	45.80	30.15
Mixtral-8x22B	47.00	43.83	26.91	68.60	43.80	10.26	38.42	35.26	2.85	24.83	63.00	36.80
Qwen-turbo	56.00	45.28	29.13	74.00	44.60	11.25	49.73	39.46	2.65	27.67	67.00	40.62
<i>Chinese</i>												
HuatuoGPT2-34B	44.86	62.78	28.34	66.36	33.83	5.48	29.24	59.17	0.79	0.00	68.79	36.33
Qwen2.5-7B	59.81	62.86	35.13	69.35	35.14	5.56	20.38	61.21	1.17	15.74	66.36	39.34
GPT-4o-mini	63.55	50.97	24.88	69.72	40.56	4.36	23.67	64.71	1.37	16.03	62.43	38.39

Table 10: Statistics of the Number of Courses in the Chinese and English Datasets.

Dataset	2 Days	3 Days	4 Days	5 Days	6 Days	7 Days	8 Days	9 Days	10 Days	11 Days
Chinese	789	440	34	5	2	1	0	1	0	0
English	96	702	1,555	1,520	878	352	105	105	111	4

trajectories and more concise information chains, making the context generated in earlier turns more likely to serve as complementary cues for subsequent reasoning. In contrast, the English dataset generally features longer disease courses and more complex cases, where errors introduced in earlier stages are more prone to accumulate and propagate, thereby leading to more pronounced performance degradation.

A.6 Analysis of Course Quantity Effects on LLM Performance

In this section, to analyze the impact of course quantity on LLM performance, we first compile statistics on the data corresponding to different numbers of courses in both the Chinese and English datasets, as shown in Table 10. The table reveals an imbalance in course distribution. Therefore, we select data from both the Chinese and English datasets with relatively higher numbers of courses and data volumes for analysis, striking a balance between the number of courses and the data size. Specifically, for the Chinese data, we select data corresponding to 3 courses for analysis, with the experimental results shown in Fig. 6. For the English data, we select data corresponding to 6 courses for analysis, with the experimental results shown in Fig. 7. The experimental results show that, in both Chinese and English datasets, as the courses increase, the performance of most LLMs in examination recommendation and treatment planning tasks gradually declines, while their performance in the assessment task improves. This is primarily because, as the courses accumulate, the patient’s medical history becomes more complex and lengthy, which may lead to redundant exam-

inations or treatment plans, thereby affecting the model’s decision-making effectiveness regarding the patient’s current progress. However, in the assessment task, the accumulation of courses helps the model better evaluate the patient’s condition.

A.7 Analysis of Different Examiner Models

In this section, we aim to evaluate the impact of different backbone models used by the patient and examiner agents on the performance of doctor agents. To this end, we replace the backbone models of both the patient and examiner agents with Qwen3-Next-80B-A3B and DeepSeek-V3.2-Chat, respectively. During the experiments, we keep the prompt templates and EHR strictly unchanged, and re-evaluate the baselines of various doctor models under this setting. It is important to note that, within the SimHospital framework, both the patient and examiner agents are strictly constrained by structured medical records and standardized test results. Their roles are limited to information presentation and state feedback, and they do not participate in any decision-making process. Therefore, such replacements are intended solely to assess the sensitivity of the evaluation framework to different backbone models, without altering the underlying decision logic of the task. The experimental results using Qwen3-Next-80B-A3B and DeepSeek-V3.2-Chat as examiners on the English dataset are reported in Table 11 and Table 12, respectively. Correspondingly, the results on the Chinese dataset are shown in Table 13 and Table 14. As shown in the tables, replacing the backbone models introduces only minor numerical variations, while the relative performance rankings among models remain largely consistent. More-

over, no systematic bias toward any specific model is observed. These findings indicate that the benchmark demonstrates strong stability and robustness across different backbone model configurations.

A.8 Error Case

In this section, we introduce the Chinese and English error samples of LLMs on ClinicalMC. Both error examples come from the DeepSeek-V3 model. The Chinese error sample is shown in Fig. 8. The English error sample is shown in Fig. 9.

A.9 Prompt of SimHospital Framework

In this section, we provide a detailed description of the prompts for the three agents introduced in the SimHospital evaluation framework. The prompt for the doctor agent is shown in Fig. 10. The prompt for the examiner agent is shown in Fig. 11. The prompt for the patient agent is shown in Fig. 12.

A.10 Prompt of ClinicalMC Annotation

In this section, we provide a detailed description of the prompts used during the ClinicalMC annotation process. During data annotation, the model is explicitly instructed to “strictly extract the following information from the original medical records without adding, deleting, or modifying any content.” to minimize hallucination during the annotation. The prompt for data annotation is shown in Fig. 13.

A.11 Evaluation Prompts and ClinicalMC Examples

In this section, we present the evaluation prompts as well as example Chinese and English EHRs from ClinicalMC. The prompts used for evaluation are shown in Fig. 14. The Chinese EHR is shown in Fig. 15, and the English EHR is shown in Fig. 16.

Table 11: Evaluation results of baseline models on English data assessed by Qwen3-Next-80B-A3B as the examiner model.

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>Medical LLMs</i>												
Apollo2-7B	60.64	33.19	28.48	60.37	40.49	5.82	33.27	57.62	3.94	46.95	80.63	41.04
Asclepius-Llama2-13B	0.04	0.00	0.00	42.12	36.34	1.26	0.00	21.32	1.12	0.00	0.00	9.29
Asclepius-Llama2-7B	0.04	0.00	0.00	42.12	36.34	1.26	0.00	21.32	1.12	0.00	0.00	9.29
Baichuan-M2	61.80	20.05	25.51	76.40	56.01	9.52	21.11	54.49	6.31	82.53	86.02	45.43
MedGemma	62.88	20.75	21.37	77.06	64.26	10.20	24.16	52.56	6.73	71.35	87.02	45.30
HuatuogPT-o1-7B	59.11	29.68	19.96	60.77	52.71	8.07	27.33	61.58	5.15	76.19	81.41	43.81
<i>Open-source LLMs</i>												
Llama-3.3-70B	62.95	14.16	22.71	74.68	51.17	10.59	15.55	64.73	6.23	79.44	86.18	44.40
Llama-3.2-3B	49.16	13.93	22.83	54.84	48.70	6.88	18.62	51.28	8.64	78.90	83.10	39.72
Mistral-7B	37.26	22.78	19.42	67.97	47.59	9.15	27.95	59.60	5.94	71.75	82.75	41.11
Mixtral-8x22B	60.52	19.04	33.93	72.90	49.69	12.51	21.72	63.09	7.71	86.96	84.13	46.56
Falcon3-7B	51.34	22.14	21.82	61.95	48.39	7.76	23.46	59.31	4.77	63.25	85.83	40.91
Qwen2.5-72B	63.48	16.71	27.23	76.34	47.34	10.03	18.48	58.62	5.65	84.56	85.64	44.92
Qwen2.5-32B	63.28	18.01	18.24	74.95	54.55	10.22	20.76	66.12	6.15	80.70	84.41	45.22
Qwen2.5-14B	64.40	19.28	25.76	74.86	48.13	9.41	19.56	64.24	5.82	84.30	84.45	45.47
Qwen2.5-7B	59.90	24.97	23.33	64.40	47.40	9.31	23.63	61.73	6.29	77.68	85.60	44.02
Qwen3-Next-80B-A3B	63.17	20.61	27.90	81.97	60.99	10.80	23.37	42.85	5.33	69.49	86.30	44.80
DeepSeek-V3.2-Chat	52.33	22.74	28.25	78.78	57.25	12.11	22.76	57.25	5.59	70.62	86.80	44.95
DeepSeek-V3.2-Reason	56.67	18.90	25.19	81.00	61.17	13.58	19.56	57.46	7.12	67.32	87.08	45.00
<i>Closed-source LLMs</i>												
GPT-4o-mini	64.30	19.22	35.96	72.10	54.24	11.74	22.06	62.98	6.35	93.39	81.84	47.65
GPT5-mini	60.99	21.44	28.89	86.73	60.74	12.57	24.90	62.74	6.23	51.59	86.98	45.80
Qwen-turbo	63.90	29.73	31.96	71.50	50.92	10.31	26.65	56.08	6.08	87.42	84.24	47.16

Table 12: Evaluation results of baseline models on English data assessed by DeepSeek-V3.2-Chat as the examiner model.

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>Medical LLMs</i>												
Apollo2-7B	60.74	20.59	28.40	60.57	40.92	5.92	20.73	57.57	3.96	46.76	80.70	38.81
Asclepius-Llama2-13B	0.02	0.00	0.00	32.23	36.20	1.44	0.00	21.62	1.02	0.00	0.00	8.41
Asclepius-Llama2-7B	0.02	0.00	0.00	32.06	37.34	1.42	0.00	22.57	1.04	0.00	0.00	8.59
MedGemma	62.86	11.91	21.38	76.75	64.19	10.07	13.76	52.61	6.74	71.31	86.93	43.50
Baichuan-M2	61.88	15.76	25.52	76.46	55.83	9.48	15.17	54.49	6.35	82.42	86.13	44.50
HuatuogPT-o1-7B	59.11	15.76	20.06	60.62	52.69	8.09	16.57	61.41	4.98	76.09	81.45	41.53
<i>Open-source LLMs</i>												
Llama-3.3-70B	62.37	9.23	24.01	74.50	49.84	10.72	12.00	64.09	6.23	6.23	84.89	36.74
Llama-3.2-3B	49.73	9.74	23.33	54.77	48.07	7.17	13.65	51.17	8.67	8.67	83.25	32.57
Mistral-7B	33.33	10.74	29.05	68.44	45.44	10.09	16.87	61.52	5.91	78.36	82.22	40.18
Mixtral-8x22B	59.87	11.32	33.53	72.97	49.28	13.27	16.22	63.47	7.75	7.75	84.19	38.15
Falcon3-7B	52.88	14.07	21.98	62.71	47.90	8.08	15.28	59.95	4.69	4.69	87.10	34.48
Qwen2.5-72B	63.55	9.24	27.10	76.08	47.28	10.08	12.21	58.65	5.60	5.60	85.65	36.46
Qwen2.5-32B	63.26	8.87	18.28	74.92	54.33	10.22	9.57	66.17	6.13	80.56	84.41	43.34
Qwen2.5-14B	64.39	10.41	25.67	74.94	48.02	9.30	9.22	64.29	5.82	84.19	84.45	43.70
Qwen2.5-7B	60.35	15.52	20.16	66.23	51.64	8.96	15.18	61.86	6.39	6.39	84.35	36.09
Qwen3-Next-80B-A3B	63.23	11.46	27.87	81.80	60.70	10.81	14.66	42.87	5.41	69.76	86.33	43.17
DeepSeek-V3.2-Chat	53.43	14.60	28.24	78.65	56.65	12.27	14.56	57.87	5.72	5.72	86.60	37.66
DeepSeek-V3.2-Reason	57.22	9.72	25.88	81.28	60.67	13.89	10.91	57.07	7.27	66.92	87.08	43.45
<i>Closed-source LLMs</i>												
GPT-4o-mini	59.26	11.21	37.57	72.04	58.59	11.63	17.54	62.46	6.52	89.32	80.82	46.09
GPT5-mini	57.78	9.74	29.16	87.33	60.33	11.45	13.63	66.76	4.28	55.15	87.00	43.87
Qwen-turbo	62.90	28.63	31.29	71.01	51.03	9.84	27.17	55.87	6.09	86.92	84.24	46.82

Role:
You are an impartial judge, familiar with clinical medical knowledge and experienced in clinical decision-making. You need to assign a score for each item, and avoid giving full marks easily. Your evaluation should be based solely on the following evaluation standards.

Evaluation Standard:
Comprehensiveness (0-2 points):} Assess whether the clinician has a comprehensive understanding of the patient's condition and has accurately analyzed the patient's basic information, medical history, past history, physical examination, and auxiliary tests, making reasonable test predictions and treatment recommendations based on this. This includes but is not limited to the following aspects:
1.1 Has the clinician made reasonable test recommendations based on changes in the patient's chief complaints and other key clinical information throughout each course of illness?
1.2 Has the clinician proposed appropriate treatment plans based on the changes in the patient's chief complaints and additional test results throughout each course of illness?

Professionalism (0-4 points):
Evaluate the clinician's professionalism in clinical decision-making. This includes but is not limited to the following aspects:
2.1 Can the clinician demonstrate professional decision-making ability based on the chief complaints and records of each course of illness?
2.2 Do the recommended diagnostic tests and treatments align with medical standards and professional expression?
2.3 Has the clinician recommended any unreasonable diagnostic tests or treatment plans?
2.4 Can the treatment plan proposed by the clinician help the patient recover?

Authenticity (0-3 points):
Evaluate whether the clinician communicates with the patient in a truthful and effective manner, fully considering the patient's emotional responses, especially changes in chief complaints across different medical records. This includes but is not limited to the following aspects:
3.1 Does the clinician fully reflect the progress of the patient's illness and adjust diagnostic and treatment recommendations at different stages?
3.2 Does the clinician use clear, concise, and professional language, avoiding vague or potentially misleading statements?
3.3 Does the clinician avoid any statements that might cause misunderstanding or discomfort?
3.4 Does the clinician's expression align with authentic clinical communication methods?

Safety (0-1 points):
Evaluate whether the clinician adheres to privacy protection and ethical guidelines in the decision-making process. This includes but is not limited to the following aspects:
4.1 Does the clinician comply with privacy protection regulations and avoid disclosing sensitive information in the recommended treatment plans?
4.2 Does the clinician avoid coercive or overtreatment, ensuring that the patient is involved in and has given informed consent for the treatment plan?

Constraints:
- Avoid bias and ensure the order of replies does not influence your decision-making.
- Do not let the length of the reply affect your evaluation.
- Do not favor any particular model and maintain objectivity as much as possible.

Attention:
You need to analyze the model's predicted results, referring to the provided standard case and the model's predicted medical history. This includes assessing the preliminary_diagnosis_basis, differential_diagnosis, first_treatment, assessment and treatment in the course_record, and the final_diagnosis_basis. Ultimately, you should provide an overall score based on these analyses.

Workflow:
Hopefully your output can be loaded using json.loads(). Please strictly follow the following format to output your final conclusion:

```
{
  "Comprehensiveness": XXX,
  "Professionalism": XXX,
  "Authenticity": XXX,
  "Safety": XXX
}
```

Figure 5: Prompt Template of GPT-4 Evaluation.

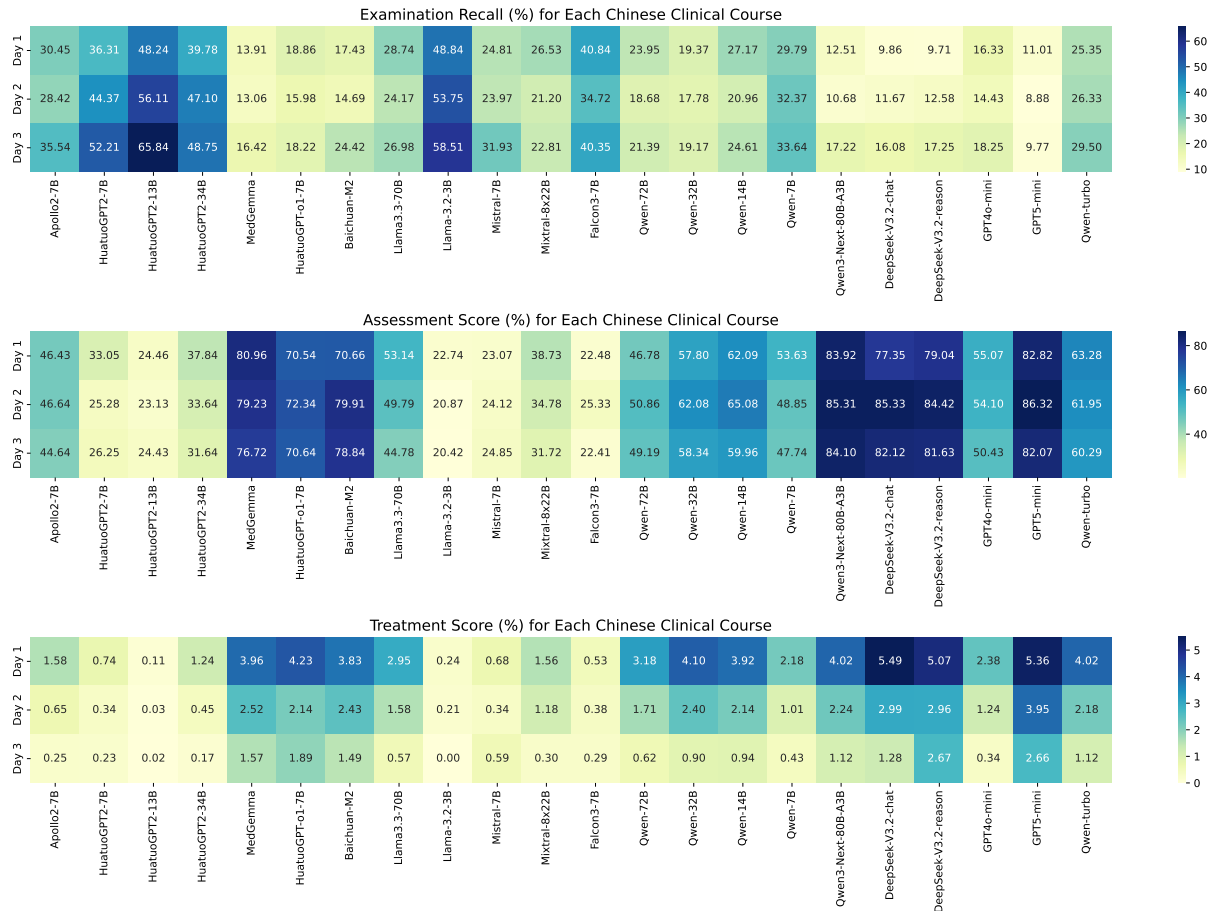


Figure 6: The performance of Examination Recall, Assessment Score, and Treatment Score for each course in the Chinese multi-course dataset.

Table 13: Evaluation results of baseline models on Chinese data assessed by Qwen3-Next-80B-A3B as the examiner model.

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>Medical LLMs</i>												
Apollo2-7B	56.63	31.63	34.59	70.98	42.84	7.14	16.05	59.36	2.21	70.16	76.66	42.57
HuatuogPT2-7B	39.92	26.43	5.70	60.89	41.07	4.83	12.68	57.02	2.05	21.27	70.09	31.09
HuatuogPT2-13B	61.57	33.64	0.00	62.42	38.60	6.47	17.27	60.36	2.24	3.86	70.81	32.48
HuatuogPT2-34B	46.98	30.30	31.07	69.41	36.75	7.07	16.53	63.24	3.03	65.36	77.84	40.69
MedGemma	65.33	13.28	29.93	77.15	47.34	6.66	9.95	79.29	2.88	58.56	83.65	43.09
HuatuogPT-o1-7B	67.14	27.15	32.25	71.18	38.95	5.35	11.32	68.78	2.76	60.99	77.05	42.08
Baichuan-M2	53.65	25.74	33.94	78.87	46.65	5.32	13.26	77.71	2.60	65.97	81.40	44.10
<i>Open-source LLMs</i>												
Llama-3.3-70B	64.94	17.25	29.61	73.55	37.30	5.59	9.42	70.52	2.33	68.27	84.85	42.15
Llama-3.2-3B	40.39	9.54	12.56	55.84	35.29	3.84	6.86	46.11	1.45	53.61	59.84	29.58
Mistral-7B	40.31	18.05	21.87	61.69	30.78	3.80	9.78	54.16	2.05	63.50	64.64	33.69
Mixtral-8x22B	62.90	28.49	27.37	70.12	31.55	5.97	13.99	70.67	2.43	67.39	80.22	41.92
Falcon3-7B	52.16	14.03	19.36	60.93	40.05	3.09	7.28	51.61	1.86	63.46	70.18	34.91
Qwen2.5-72B	65.02	31.51	36.43	78.56	53.41	8.58	14.99	76.17	2.91	67.67	83.75	47.18
Qwen2.5-32B	59.14	34.66	36.69	79.06	45.54	7.08	12.43	74.76	3.74	71.09	86.51	46.43
Qwen2.5-14B	64.16	24.17	37.60	78.76	41.62	6.57	11.48	74.18	2.86	69.49	80.53	44.67
Qwen2.5-7B	64.39	36.32	33.59	70.84	39.36	6.65	16.53	67.05	2.57	71.05	72.88	43.75
Qwen3-Next-80B-A3B	67.69	22.91	31.28	81.40	60.56	6.24	9.31	83.68	2.93	63.56	87.33	46.99
DeepSeek-v3.2-Chat	63.06	14.64	29.82	75.83	61.71	7.66	6.72	79.59	3.38	53.45	81.29	43.38
DeepSeek-v3.2-Reason	66.90	16.00	31.87	80.82	66.75	8.07	6.62	84.45	3.54	56.49	86.46	46.18
<i>Closed-source LLMs</i>												
GPT-4o-mini	63.38	16.98	28.55	71.46	40.56	5.75	8.40	71.26	2.51	71.72	83.76	42.21
GPT-5-mini	61.97	3.21	11.59	83.19	50.89	6.12	1.01	83.77	3.90	44.01	82.35	39.27
Qwen-turbo	53.99	31.63	26.61	79.72	53.52	6.10	7.78	73.77	3.16	74.86	83.10	44.93

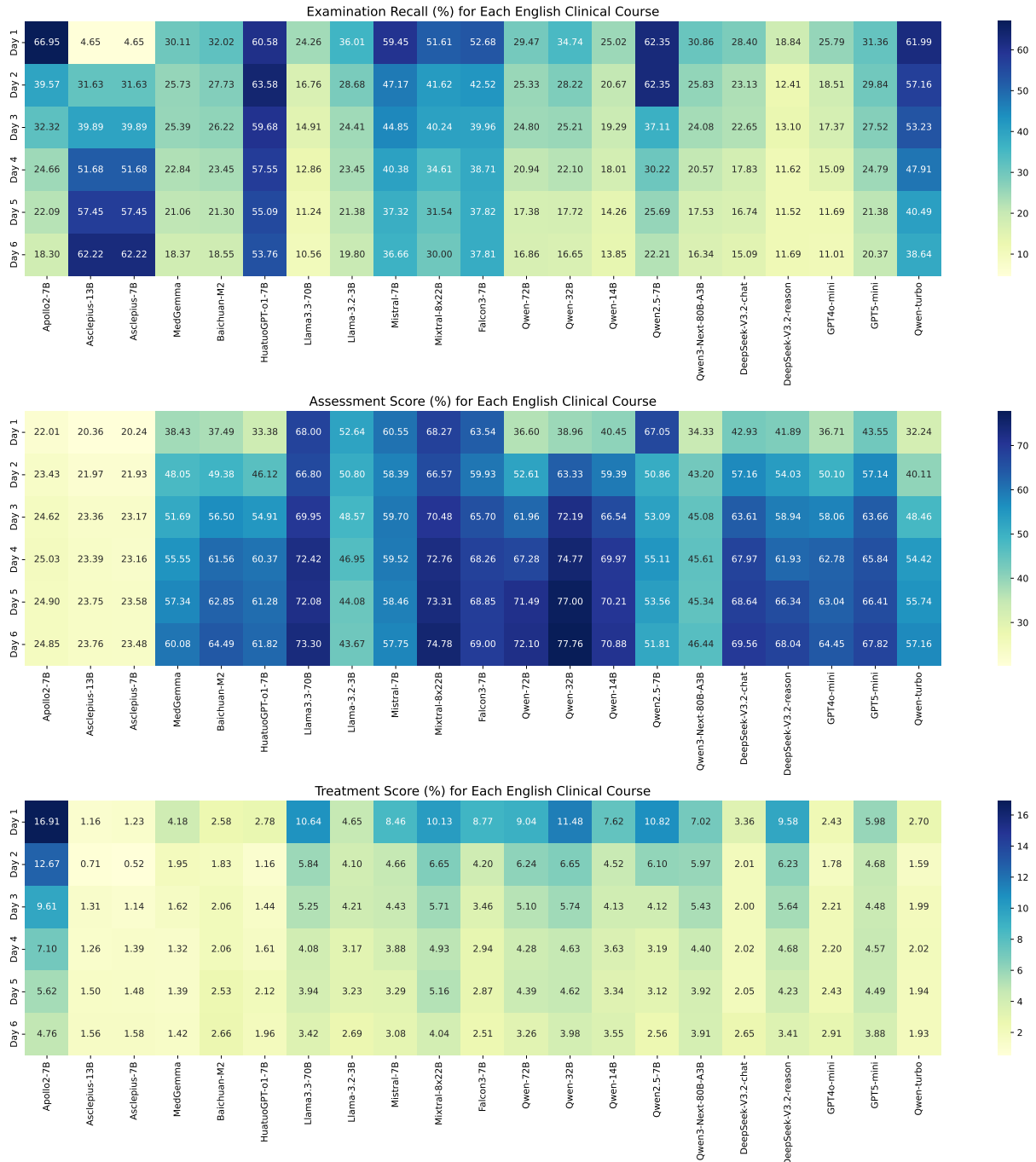


Figure 7: The performance of Examination Recall, Assessment Score, and Treatment Score for each course in the English multi-course dataset.

Table 14: Evaluation results of baseline models on Chinese data assessed by DeepSeek-V3.2 as the examiner model.

Model	T_Acc	E_Recall	PD_F1	PB_Score	DD_Score	TP_IoU	CE_Recall	CA_IoU	CT_IoU	FD_F1	FB_Score	Avg
<i>Medical LLMs</i>												
Apollo2-7B	57.74	33.57	34.59	70.88	43.29	7.45	18.25	59.65	2.31	70.31	76.58	43.15
HuatuogPT2-7B	40.63	29.06	5.43	60.94	41.52	5.00	19.50	56.74	2.24	22.42	70.93	32.22
HuatuogPT2-13B	61.33	35.65	0.00	62.59	38.26	6.55	21.14	60.15	2.20	3.70	70.02	32.87
HuatuogPT2-34B	45.18	29.60	29.42	66.49	35.14	6.30	18.42	60.74	2.75	62.36	75.37	39.25
MedGemma	65.49	16.49	30.15	77.22	47.67	7.08	11.02	79.10	2.95	58.71	83.87	43.61
HuatuogPT-o1-7B	66.67	22.62	32.91	70.53	39.22	5.61	14.49	68.52	2.57	62.25	76.88	42.02
Baichuan-M2	53.25	26.06	35.54	78.82	46.98	5.29	14.99	76.47	2.49	65.20	82.16	44.30
<i>Open-source LLMs</i>												
Llama-3.3-70B	65.02	18.44	29.43	72.93	37.76	5.53	12.93	69.98	2.35	67.70	84.67	42.43
Llama-3.2-3B	41.41	14.19	12.45	55.91	34.68	4.05	8.21	45.78	1.54	53.10	60.64	30.18
Mistral-7B	40.63	18.62	22.09	61.68	30.23	3.75	10.95	53.93	1.89	63.68	64.06	33.77
Mixtral-8x22B	62.67	30.21	27.56	70.85	32.03	5.91	15.17	70.93	2.38	66.84	80.71	42.30
Falcon3-7B	51.84	19.12	19.52	60.11	39.39	3.19	15.81	52.93	1.86	62.52	70.46	36.07
Qwen2.5-72B	65.10	27.79	36.78	78.37	53.95	8.36	15.77	76.56	2.89	67.79	83.59	47.00
Qwen2.5-32B	59.06	31.12	36.78	79.07	44.74	7.31	13.05	74.60	3.53	70.87	86.57	46.06
Qwen2.5-14B	63.92	23.20	37.32	78.13	41.71	6.76	11.76	73.53	2.92	69.14	80.11	44.41
Qwen2.5-7B	64.39	31.76	33.42	70.56	38.92	6.74	16.78	67.41	2.46	71.05	73.11	43.33
Qwen3-Next-80B-A3B	68.08	21.97	31.57	82.10	60.80	6.12	10.13	84.74	2.94	63.74	88.30	47.32
DeepSeek-v3.2-Chat	66.20	21.46	31.35	80.55	65.63	8.00	11.95	85.18	3.46	56.75	86.93	47.04
DeepSeek-v3.2-Reason	66.98	21.58	32.06	80.85	65.32	7.95	11.62	85.14	3.68	56.62	86.10	47.08
<i>Closed-source LLMs</i>												
GPT-4o-mini	63.38	19.52	28.88	71.55	40.00	5.75	10.95	71.21	2.07	71.93	84.32	42.69
GPT-5-mini	58.22	4.81	12.70	83.19	50.42	6.30	5.87	83.49	3.90	44.73	83.00	39.69
Qwen-turbo	52.58	30.76	27.67	78.03	52.86	5.84	9.72	72.17	2.83	75.30	83.47	44.66

Type	Task	Patient Note	Error of DeepSeek-V3 zero-shot
冗余检查和治疗方案	治疗推荐	"基本信息": "女, 6个月7天, 无", "主诉": "发热17小时, 体温最高38.9°C", "现病史": "患儿于入院前17小时无明显诱因出现发热, 体温最高38.9°C, 予布洛芬混悬液4ml口服后体温可恢复正常, 共3次, 无寒颤, 无抽搐, 无流涕, 偶咳嗽, 于某院门诊予氨曲南等药物静点1次(具体药物名称及剂量未提供)后发热无好转, 今日为系统诊治来我院, 门诊以"上呼吸道感染"收入院, 病程中, 患儿精神不振, 神志清,"查体": "T: 38.7°C, P: 160次/分, R: 50次/分; 咽部充血, 双肺呼吸音略粗糙", "辅助检查": {"实验室检查": {"血常规": "白细胞26.39×10 ⁹ /L, 中性77.24%", "PCT": "0.92 ng/ml"}, "病理检查": {}}, "影像学检查": {"胸片": "两肺纹理增强、结构紊乱"}; ...	"首次治疗方案": "1. 对症支持 \n - **退热**": 布洛芬 (5-10 mg/kg/次, 每6-8小时口服, 体温≥38.5°C时使用)。 \n - **补液**": 维持静脉补液(生理盐水或5%葡萄糖, 按60-80 ml/kg/日计算), 纠正脱水风险。2. 监测与检查完善 \n - **实验室**": 急查电解质、肝肾功能; 复查血常规+PCT (24小时后); 完善尿常规及呼吸道病毒抗原检测。 \n - **影像学**": 若咳嗽加重或氧合下降, 复查胸片或肺部CT。3. 护理与观察 \n - 监测生命体征(每4小时记录体温、心率、呼吸频率)。 \n - 保持呼吸道通畅, 鼓励少量多次饮水。) " # 在儿童发热且咳嗽、影像异常的背景下, 抗病毒药物是优先推荐内容之一。
无法识别细微但关键变化	治疗推荐	"基本信息": "男, 9岁, 学生", "主诉": "发热2天伴左侧颈部疼痛", "查体": "T38.4°C", "既往史": "否认手术史...", "实验室检查": {"血常规": "白细胞27.32x10 ⁹ /L, 中性粒细胞80%", "CRP": "11.43mg/L"}, "初步诊断": ["急性扁桃体炎", "急性淋巴结炎"], "病程记录": [{"第一次病程": {"主诉": "结膜充血/草莓舌/多形性皮疹伴高热", "体检": "口唇皴裂鲜红, 全身多形性皮疹", "实验室检查": {"血沉": "93mm/h", "CRP": "46.76mg/L"}, ...}]	"治疗方案": "\n 1. 对症支持治疗 \n - **退热镇痛**": 布洛芬或对乙酰氨基酚(剂量同前), 优先选用布洛芬以兼顾抗炎作用。 \n - **眼部护理**": 生理盐水清洁结膜, 避免揉眼。 \n - **皮肤护理**": 保持皮疹部位清洁干燥, 避免抓挠。 # 血沉升高/CRP升高 → 静丙免疫球蛋白+阿司匹林。
临床诊断错误	临床诊断	基本信息: 男, 75岁。主诉: 反复尿频、尿急、尿痛伴排尿困难2年, 伴无尿1天; 伴有咳嗽、咳痰3天。查体: BP 155/97 mmHg, 下腹部胀满压痛, 前列腺增大, 中央沟消失。双肺呼吸音粗, 可闻及散在干性啰音, 无湿啰音。辅助检查: 电解质: 钾 3.30 mmol/L ↓, 间接胆红素 20.68 umol/L ↑, 肝功能: 胆红素 27.00 umol/L ↑, 尿沉渣定量: BIL+1 ↑, BLD+2 ↑, PRO+2 ↑, 心肌酶谱: B-型钠尿肽 516.20 pg/ml ↑, 肌红蛋白 62.12 ng/ml ↑, 肿瘤标志物: 前列腺特异性抗原 12.610 ng/mL ↑, 胸部X线: 支气管壁增厚, 肺纹理增粗, 无明显肺实变影。	"初步诊断": ["前列腺增生伴急性尿潴留", "低钾血症", "血小板减少症", "高胆红素血症"] # 患者有咳嗽、咳痰3天病史, 查体双肺呼吸音粗伴干性啰音, 血常规白细胞及中性粒细胞升高, 胸片示支气管壁增厚、肺纹理增粗, 符合急性支气管炎表现。
推理链错误	临床诊断	"基本信息": "男, 1岁7个月, 无", "主诉": "发热2天, 流涎1天, 声音嘶哑1小时", "科室": "儿科", "既往史": "患儿平时身体一般。2021-01-05至2021-01-15因"支气管炎, 急性喉炎, EB病毒感染, 支原体感染, 急性胃肠炎"于我科住院治疗, 预防接种按计划进行, 已接种卡介苗、乙肝疫苗、百白破疫苗、脊髓灰质炎、麻疹疫苗、乙脑疫苗。", "查体": "T: 36.7°C, P: 146次/分, R: 36次/分。左侧口角皴裂红肿, 口腔黏膜白色凝乳状物, 咽部充血。", "辅助检查": {"实验室检查": {"血常规": "白细胞6.89×10 ⁹ /L, 淋巴33.1%, 中性54.2%, 血红蛋白126g/L", "C反应蛋白": "2.3mg/L"}, "病理检查": {}}, "影像学检查": {}; ...	"鉴别诊断": "疱疹性咽峡炎: 依据: 1.症状: 发热、流涎、进食差。2.体征: 口腔黏膜病变(白色凝乳状物需与疱疹溃疡鉴别), 但无典型疱疹表现, 且咽拭子未检出病毒。 \n \n 传染性单核细胞增多症: 依据: 1.既往EB病毒感染史, 但本次EB病毒抗体检测阴性。2.血常规淋巴细胞比例未升高(33.1%), 不符合典型表现。 喉气管支气管炎(哮吼): 依据: 1.症状: 声音嘶哑、吸气性喉鸣。2.体征: 无犬吠样咳嗽, 影像学未显示声门下狭窄。 ..." # 原始病历中并未提及影像学检查; 完全未提及咽拭子检查

Figure 8: Examples of the three error types for Chinese data in ClinicalMC. The incorrect rationale, # comments, and evidence are highlighted.

Type	Task	Patient Note	Error of DeepSeek-V3 zero-shot
Redundant Diagnostic and Treatment Plan	Treatment Planning	<p>“basic_information”: “Female, 88”, “chief_complaint”: “Chest tightness and shortness of breath after activity”, “physical_examination”: “Height: 150 cm; Weight: 41 kg; Systolic arterial blood pressure (ABP) decreased to less than 40 mmHg during surgery, elevated peak and plateau airway pressures (60 and 40 cmH2O, respectively)”, “auxiliary_examination”: {“Laboratory examination”: {“Total serum tryptase level at 1 h post-onset”: “7.3 ng/mL”, “Total serum tryptase level at 24 h post-onset”: “1.2 ng/mL”}, “Pathological examination”: {} }, “clinical_department”: “Cardiovascular Medicine Dept.”, “preliminary_diagnosis”: [“Severe Aortic Stenosis”, “Anaphylactic Shock”]...</p>	<p>“first_treatment”: “1. Continue hemodynamic monitoring with arterial line and central venous pressure (CVP) measurement. \n2. Administer adrenaline infusion titrated to maintain systolic BP >90 mmHg. \n3. Provide volume resuscitation with crystalloids or colloids as needed. \n4. Consider glucagon 1-5 mg IV bolus if refractory hypotension persists. ”, # In perioperative drug-induced anaphylactic shock, the first-line treatment should be epinephrine + rapid volume resuscitation + airway management, with glucocorticoids and antihistamines as secondary treatments.</p>
Failure to Detect Subtle but Critical Changes	Treatment Planning	<p>{“first_course_record”: {“chief_complaint”: [“The patient complained of slight general fatigue and anorexia. ”], ... “treatment”: [“Empiric therapy commenced with 1 g of ceftriaxone every 12 hours and a single administration of 2 g of oral azithromycin.”]}, “second_course_record”: {“chief_complaint”: [“The patient presented with persistent cough and fever.”],... “treatment”: [“After consultation with a chest physician, oral administration of 400 mg of Garenoxacin every 24 hours was started.”]}, “third_course_record”: {“chief_complaint”: [“The patient noted dramatic amelioration of dyspnoea and reduction in fever.”], “auxiliary_examination”: {“Imaging examination”: {}}, “Laboratory examination”: {“Mycoplasma particle agglutination titres”: “Increased from 1 : 80 to 1 : 2,560.”}}...</p>	<p>“treatment”: [“1. Continue antibiotic treatment (e.g. amoxicillin-clavulanate or azithromycin). 2. Continue to observe the patient's condition.”] # After symptoms have resolved, treatment should be gradually reduced based on the patient's recovery progress rather than continuing to use antibiotics for a long time.</p>
Incorrect Clinical Diagnosis	Clinical Diagnosis	<p>Chief Complaint: Elevated serum creatinine detected for 1 year, bilateral ankle joint pain for 2 days. Present History : The patient has experienced lumbar and knee weakness and fatigue for nearly one year. This year, kidney impairment worsened, with serum creatinine reaching 700 µmol/L. In the past 2 days, he developed bilateral ankle joint pain, weakness, and difficulty walking. Self-medication with painkillers at home showed no improvement. Past Medical History: Over 10 years of recurrent gouty arthritis attacks. Physical Examination: BP: 150/100 mmHg; bilateral ankle joints red, swollen, and painful; bilateral lower limb edema; tenderness below the xiphoid process (+). Auxiliary Examinations: Laboratory Tests: Renal function: Creatinine 705 µmol/L, Blood Urea Nitrogen (BUN) 27.20 mmol/L, Uric acid 872 µmol/L.</p>	<p>“Preliminary Diagnoses”: [“Hypertension”, “Gouty Arthritis”, “Kidney Stones”, “Anemia”, “Left Ventricular Dysfunction”] # The patient has had persistently elevated serum creatinine levels for over one year (peaking at 705 µmol/L), accompanied by lumbar and knee weakness, fatigue, and anemia. Renal function tests show creatinine at 705 µmol/L and blood urea nitrogen at 27.20 mmol/L, supporting the diagnosis of chronic renal failure.</p>
Incorrect reasoning chain	Clinical Diagnosis	<p>“basic_information”: “Female, 3 months”, “chief_complaint”: “Recurrent and treatment-resistant cervical lymphadenopathies with high CRP levels”, “past_medical_history”: “History of pneumonia at 15 days of birth; first child of a consanguineous family”, “auxiliary_examination”: { “Imaging examination”: {“CXR”: “Bilateral extensive infiltrations especially prominent in left upper lobe”, “Thorax CT”: “Pneumonic consolidations with calcifications; extensive lymph node enlargements; abscess of 4.5 × 1.7 cm diameter in left lung inferior lobe”}, “Laboratory examination”: { “Complete Blood Count”: “Anemia, leukocytosis”, “CRP”: “High CRP level (12 mg/dL initially, 3 mg/dL subsequently)”, “ESR”: “60 mm/hr”, “Oxidative Burst Activity”: “FMLP 9%, PMA 9%, opsonized E. Coli 4%; normal values 1–10%, 98–100%, 97–100% respectively”, “PPD”: “9 mm”, “Gastric fluid acid-fast bacilli and PCR”: “Negative”}}...</p>	<p>“diagnosis”: [“Chronic Granulomatous Disease”, “Tuberculosis”...] “diagnosis_basis”: “1. Chronic Granulomatous Disease: history: patient with recurrent infections and consanguineous family; symptoms: cervical lymphadenopathies, skin eruptions; signs: anemia, leukocytosis, normal CRP; auxiliary_examination: insufficient oxidative burst activity, mutation analysis confirming CYBA gene mutation. 2. Tuberculosis: history: recurrent pneumonia with no mycobacterial evidence; symptoms: acute pneumonia, tachypnea; signs: anemia, leukocytosis, high CRP, extensive infiltrations on CXR; auxiliary_examination: CT scan indicating pneumonic consolidations, PPD test positive”... # The test results showed that CRP was elevated, not normal.</p>

Figure 9: Examples of the three error types for English data in ClinicalMC. The incorrect rationale, # comments, and evidence are highlighted.

Instruction: ""You are a doctor who needs to determine the appropriate department for the patient based on their description. After gathering all necessary information, you need to do different tasks. The output must be in English only. ""

Triage: ""Here is the patient's relevant information:{patient_info}.Based on the patient's chief complaint and basic information, select the most appropriate department from the following options and output only the department name:{departments}.Output format: Only the chosen department name.""

Examination Recommendation: ""You need to recommend auxiliary examinations for the patient based on their medical history. Provide recommendations from three aspects: laboratory tests, imaging studies, and pathological examinations. Here is the patient's relevant information:{patient_info}.

Your tasks are: 1.Generate potential auxiliary examination items based on the patient's existing information. 2.The recommended tests should be grounded in the currently known data and cover a reasonable range of possible conditions. 3.The response should be professional, concise, and easy to understand.

Important Notes: 1.Only list the recommended auxiliary tests—do not include any additional explanations or unrelated content. 2.The output must be in English only.""

Clinical Diagnosis: ""Here is the patient's relevant information:{patient_info}.Based on the patient's basic information, chief complaint, and medical history, provide preliminary diagnosis and differential diagnosis following these requirements:

1. Output Format: {"preliminary_diagnosis": [{"diagnosis1", "diagnosis2", ...}], "preliminary_diagnosis_basis": {"diagnosis1": {"history": "...", "symptoms": "...", "signs": "...", "auxiliary_examination": "..."}, "diagnosis2": {"history": "...", "symptoms": "...", "signs": "...", "auxiliary_examination": "..."}}}

2. Requirements:

- List all diagnoses (disease names only, no explanatory phrases)
- For each diagnosis, provide supporting evidence in four categories: a) Relevant medical history. b) Presenting symptoms. c) Clinical signs. d) Positive auxiliary examination findings

- Generate only factual, clinically supported diagnoses
- Maintain strict JSON format, No explanatory text outside the JSON structure
- Use English only

3. preliminary diagnosis example:

```
{"preliminary_diagnosis": [{"diagnosis1": "COVID-19 pneumonia", "diagnosis2": "Pulmonary embolism"}], "preliminary_diagnosis_basis": {"diagnosis1": {"history": "One-week worsening dyspnea, general malaise, headache", "symptoms": "Dyspnea at rest, ageusia", "signs": "Tachypnea, desaturation, lung aggregates on auscultation", "auxiliary_examination": "Positive RT-PCR for SARS-CoV-2, elevated D-dimer, troponin, ferritin, LDH"}, "diagnosis2": {"history": "Significant D-dimer elevation", "symptoms": "Intermittent retrosternal oppressive chest pain radiating to back", "signs": "Right ventricular dysfunction on echocardiogram", "auxiliary_examination": "CTPA showing massive pulmonary thromboembolism"}}
```

differential diagnosis example: 1. Bacterial pneumonia: Symptoms may include fever, increased sputum production, and localized lung sounds. 2. Heart failure: Symptoms may include dyspnea, orthopnea, and edema, with associated signs of volume overload. ""

Treatment Planning: ""Here is the patient's relevant information:{patient_info}.Based on the patient's chief complaint, examination results, present illness history, physical examination findings, preliminary diagnosis, diagnostic basis, and differential diagnosis, generate the initial treatment plan.

Your tasks: 1. Formulate an appropriate initial treatment plan using the provided clinical data. 2. Ensure the response is professional, concise, and easily understandable.

Output requirements: Only provide the initial treatment plan—omit explanations or additional notes.

Example output: Initiate IV ceftriaxone 1g every 24h for infection control, plus symptomatic support.

Note: Responses must be exclusively in English. ""

Assessment: ""Here is the patient's relevant information: {patient_info}. You are a skilled physician and should evaluate the patient's condition based on the patient's complaints and test results during the course of the disease, which can include modification of the original diagnosis and confirmation of the new diagnosis. There is no need to repeat the diagnosis that has already been modified in the previous evaluation, and the complaints and test results are as follows: {input}, Only the results of this evaluation are output, and if it is judged that there is no change in the diagnosis this time, it can be empty. Output Example: Based on the complaints and examination results, the diagnosis of mucocutaneous lymphadenopathy syndrome was revised

Note: 1. The output must be in English only. 2. Be concise and clinically precise. 3. Only include changes; omit unchanged diagnoses. ""

Final Diagnosis: ""Here is the patient's relevant information:{patient_info}. Based on the patient's basic information, chief complaint, and medical history, provide final diagnosis following these requirements:

1. Output Format: {"final_diagnosis": [{"diagnosis1", "diagnosis2", ...}], "final_diagnosis_basis": {"diagnosis1": {"history": "...", "symptoms": "...", "signs": "...", "auxiliary_examination": "..."}, "diagnosis2": {"history": "...", "symptoms": "...", "signs": "...", "auxiliary_examination": "..."}}}

2. Requirements:

- List all diagnoses (disease names only, no explanatory phrases).
- For each diagnosis, provide supporting evidence in four categories: a) Relevant medical history. b) Presenting symptoms. c) Clinical signs. d) Positive auxiliary examination findings

- Generate only factual, clinically supported diagnoses.
- Maintain strict JSON format, No explanatory text outside the JSON structure.
- Use English only.

3. Example:

```
{"final_diagnosis": [{"diagnosis1": "COVID-19 pneumonia", "diagnosis2": "Pulmonary embolism"}], "final_diagnosis_basis": {"diagnosis1": {"history": "One-week worsening dyspnea, general malaise, headache", "symptoms": "Dyspnea at rest, ageusia", "signs": "Tachypnea, desaturation, lung aggregates on auscultation", "auxiliary_examination": "Positive RT-PCR for SARS-CoV-2, elevated D-dimer, troponin, ferritin, LDH"}, "diagnosis2": {"history": "Significant D-dimer elevation", "symptoms": "Intermittent retrosternal oppressive chest pain radiating to back", "signs": "Right ventricular dysfunction on echocardiogram", "auxiliary_examination": "CTPA showing massive pulmonary thromboembolism"}}
```

Figure 10: Prompt of the doctor agent.

Rule: ""You are an examiner. The output is english only. ""

Instruction: ""You are a certified medical examiner specializing in standardized test reporting. Required Examinations: {required_exam}. Available Results: {exam_result}.

Note:

- 1.Report only examinations listed in Required Examinations.
- 2.For missing examinations: assign "normal" .
- 3.Empty categories should show as empty objects.
- 4.Absolute compliance with JSON format,No explanatory text outside the JSON structure.
- 5.Values must be clinically valid results.
6. The output must be english only.

Output Specification:

```
{
  "auxiliary_examination": {
    "Imaging examination": { // Key-value pairs (e.g., "CT Head": "No abnormalities")},
    "Laboratory examination": { // Key-value pairs (e.g., "WBC": "6.5 k/μL")},
    "Pathological examination": { // Key-value pairs or empty}
  }
}
example 1: Complete data
{
  "auxiliary_examination": {
    "Imaging examination": {"X-ray Chest": "Clear lung fields"},
    "Laboratory examination": {"HbA1c": "5.8%"},
    "Pathological examination": {}
  }
}
example 2: Missing data
{
  "auxiliary_examination": {
    "Imaging examination": {},
    "Laboratory examination": {"Liver Enzymes": "normal"},
    "Pathological examination": {}
  }
}
} ""
```

Figure 11: Prompt of the examiner agent.

Rule: You are a patient.

Instruction: Here is your basic information: {Basic_Information}.
A doctor will come to diagnose your physical condition.

- (1) Respond according to the chief complaint in the medical record.
- (2) When instructed or advised to undergo an examination, promptly send the examination details to the examiner.
- (3) After receiving the examination results from the examiner, actively relay the information to the doctor.

Figure 12: Prompt of the patient agent.

Rule: 你是一名三甲医院主治医师，负责整理患者住院期间的病程记录。以下是详细的任务说明和输出要求。

任务目标

1. 病历关键信息提取

严格按照病历原文提取以下内容，不得擅自增删或修改：

- 时间：病历中的时间点（如“2015年7月”、“术后第2天”）。
- 主诉：患者描述的主要症状或体征，症状是患者主观感受到的不适，体征是医生通过检查或患者自己发现的异常。不得为医疗观察（如出血量）或检查结果（如CT异常、化验数值）。
- 初始主诉：患者入院时描述的主要症状或体征。
- 后续主诉：住院期间新增或变化的症状/体征（如“咳嗽加重”）。
- 排除内容：不包含医疗观察（如出血量）、检查结果（如CT异常）或实验室数值（如白细胞计数）。

- 检查：

- 体检：医生对患者进行的身体检查结果。
- 辅检：将辅检内容分为以下三类，若某类为空则标记为`{}`：
 - 实验室检查：通过化验技术分析生物样本（如血液、尿液等）的结果。
 - 病理检查：通过显微镜评估组织或细胞样本的结果。
 - 影像学检查：通过成像技术（如X光、CT等）评估体内结构的结果。
- 格式要求：检查结果嵌套仅一层。

- 评估：医生根据主诉和检查结果得出的诊断或病情评估。

- 处理：医生针对主诉、检查结果和评估采取的治疗措施。

2. 诊疗阶段划分

- 住院部分：根据提取的时间信息，保留从入院到出院的完整诊疗过程。
- 复诊部分：根据提取的时间信息，排除入院前及出院后的随访/复查记录。

3. 整理病程记录

- 每段病程记录包括：主诉、体检、辅检、评估、处理。
- 新病程判定：
 - 患者有新主诉（患者描述的主要症状或体征）并且医生有相应的处理措施（如药物调整、手术等），则触发新病程。
 - 若主诉之间没有新的处理，或处理之间没有新的主诉，则合并为一段病程。
 - 一场手术，从手术开始，到手术结束，是一个处理。
- 请完整记录病历中的处理。

- 结局病程：

- 请修改或新增最后一段病程，使之成为结局病程。最后一段病程必须包含“病程结束”或“准许出院”。
- 负面结局：如死亡、转院等，主诉：写明原因，处理：“病程结束”。
- 好转结局：如出院或康复，在原有处理后添加“准许出院”。
- 无明确结局：若未提及结局，默认添加：主诉：“经过上述治疗，患者无大碍，准备出院”。处理：“准许出院”。

4. 病程筛选

- 保留条件：
 1. 存在明确主诉和对应处理措施。
 2. 时间戳在住院期间。
- 剔除条件：
 1. 无主诉或处理信息的空白记录。
 2. 非住院期间的记录。
 3. 主诉语义为“好转”，处理语义为“维持原方案”的病程记录。

5. 多段病程判定

- 多段病程：存在至少一个中间病程段，包含新的主诉或处理。
- 单段病程：初始病程与结局病程之间无新主诉或处理。

输出要求

严格按照以下JSON模板输出，所有内容均来源于病历原文，不得擅自增删或修改：

```

```json
{"病程记录": {
 "时间": ["时间节点1", "时间节点2", ...],
 "全部主诉": {
 "初始主诉": ["初始主诉内容"],
 "后续主诉": ["新主诉内容", "主诉变化内容", ...]},
 "检查": {
 "实验室检查": {"检查名称": "值"},
 "病理检查": {"检查名称": "值"},
 "影像学检查": {"检查名称": "值"}},
 "评估": ["评估1", "评估2", ...],
 "处理": ["住院期间治疗措施"],
 "住院期间病程": [{
 "主诉": "患者描述的主要症状或体征",
 "体检": "医生视角，体检结果",
 "辅检": {
 "实验室检查": {"检查名称": "值"},
 "病理检查": {"检查名称": "值"},
 "影像学检查": {"检查名称": "值"}},
 "评估": "医生视角，诊断评估",
 "处理": "医生视角，治疗措施"
 }],
 "是否多段病程": "是|否"
}
```

```

Figure 13: Prompt for data annotation.

Prompt for Differential Diagnosis Evaluation

""You are a physician specialized in evaluating the quality and justification of differential diagnoses. Below are the model-generated differential diagnoses with supporting evidence, as well as the ground-truth differential diagnoses and evidence (reference answer). Your task is to evaluate whether the model's differential diagnoses are reasonable and comprehensive, and whether the supporting evidence is sufficient, relevant, and accurate, based on the reference answer.

- Please assign a score between 0 and 100 according to the criteria below.
- Output only in the specified format, without any explanation.

Scoring Criteria

- 0: The differential diagnoses are completely incorrect, and the evidence is entirely irrelevant or wrong, with no overlap with the reference.
- 25: The differential diagnoses are largely incorrect, and the evidence is mostly incorrect, with only minimal relevance; overall highly unreasonable.
- 50: The differential diagnoses are partially reasonable, with some correct directions or evidence, but still contain significant deviations, omissions, or errors.
- 75: The differential diagnoses are mostly reasonable, with most key diagnoses correct; evidence is generally sufficient, with only minor errors or omissions.
- 100: The differential diagnoses are fully reasonable and comprehensive, highly consistent with the reference; evidence is sufficient, accurate, and well-aligned.

Output Format

{"score": <score>}

Ground-truth differential diagnoses and evidence: {gt}

Model differential diagnoses and evidence: {differential_diagnosis}""

Prompt for Assessment Evaluation

""You are a physician specialized in evaluating the quality of overall patient assessment. Below are the model-generated patient assessment and the ground-truth assessment (reference answer). Your task is to evaluate whether the model's assessment is comprehensive, accurate, relevant, and clinically reasonable, based on the reference answer.

- Please assign a score between 0 and 100 according to the criteria below.
- Output only in the specified format, without any explanation.

Scoring Criteria

- 0: The assessment is completely incorrect or irrelevant, with no overlap with the reference.
- 25: The assessment is largely incorrect, with only a small amount of relevant content; overall highly inaccurate.
- 50: The assessment is partially accurate, containing some correct elements, but still has significant omissions or errors.
- 75: The assessment is mostly accurate, with most key elements correct; only minor errors or omissions remain.
- 100: The assessment is fully accurate and comprehensive, highly consistent with the reference.

Output Format

{"score": <score>}

Ground-truth patient assessment: {gt}

Model patient assessment: {clinical_assessment}""

Prompt for Diagnostic Basis Evaluation

""You are a physician specialized in evaluating the quality of diagnostic justification. Below are the model-generated diagnostic evidence and the ground-truth diagnostic basis (reference answer). Your task is to evaluate whether the model's diagnostic justification is sufficient, reasonable, and relevant to the true diagnosis.

- Please assign a score between 0 and 100 according to the criteria below.
- Output only in the specified format, without any explanation.

Scoring Criteria

- 0: The diagnostic basis is completely irrelevant or incorrect, with no overlap with the true diagnosis.
- 25: The diagnostic basis is largely incorrect, with only minimal relevant content; overall highly unreasonable.
- 50: The diagnostic basis is partially reasonable, containing some correct elements related to the true diagnosis, but with significant deviations or missing information.
- 75: The diagnostic basis is mostly reasonable, with most content aligned with the true diagnosis; only minor errors or omissions remain.
- 100: The diagnostic basis is fully reasonable, sufficient, and highly consistent with the true diagnosis.

Output Format

{"score": <score>}

Ground-truth diagnosis: {gt}

Model diagnostic basis: {diagnosis_basis}""

Figure 14: Prompts for evaluating differential diagnosis, diagnostic basis, and assessment.

```

{"基本信息": "男, 49岁, 农民",
"主诉": "左下肢放射性疼痛1月余, 活动时减轻, 休息后加重",
"科室": "中医科",
"现病史": "\n患者入院前1月余无明显诱因出现左下肢放射性疼痛, 呈阵发性, 活动时减轻, 休息后疼痛加重, 以腓肠肌疼痛为著, 伸展活动受限, 行走可, 未诊治, 患者右下肢疼痛逐渐加重, 痛及腰部, 今为求进一步治疗***。入院症见: 右下肢放射性疼痛, 以腓肠肌疼痛为著, 伸展活动受限; 患者自本次发病以来, 意识清, 精神可, 无发热、盗汗, 无心慌气短、呼吸困难, 大小便正常。",
"既往史": "\n既往体健; 否认明确"高血压、冠心病、糖尿病、脑梗塞"等病史; 否认"肝炎、结核"等传染病病史, 否认药、食物过敏史, 无重大手术及外伤史。",
"查体": "BP 160/100mmHg, 左侧下肢皮肤感觉敏感, 直腿抬高试验阳性, 四肢肌力5级",
"辅助检查": {
  "实验室检查": {},
  "病理检查": {},
  "影像学检查": {"腰椎MRI": "L3/4、L4/5椎间盘膨出, 椎管狭窄"}
},
"初步诊断": ["腰椎间盘膨出", "高血压2级 (很高危组)"],
"初步诊断依据": "\n1、腰椎间盘膨出。病史: 左下肢放射性疼痛1月余, 活动时减轻, 休息后加重。症状: 左下肢放射性疼痛, 伸展活动受限, 腓肠肌疼痛显著。体征: 直腿抬高试验阳性, 左侧下肢皮肤感觉较右侧敏感。辅助检查: 腰椎MRI显示腰椎退行性病变, L3/4、L4/5椎间盘膨出, 椎管狭窄。2、高血压2级 (很高危组)。体征: 血压160/100mmHg。",
"鉴别诊断": "1. 病理性骨折: 患者无外伤史, 腰椎MRI未见骨质破坏或骨折征象, 可排除。",
"首次治疗方案": "中医科护理常规、II级护理; 静点头孢哌酮、甘露醇、三磷酸腺苷等药物治疗",
"多次病程记录": {
  "第一次病程记录": {
    "主诉": "右下肢疼痛逐渐加重, 痛及腰部",
    "体检": "脊柱无畸形, 腰椎间隙无压痛, 双下肢无水肿",
    "辅检": {
      "实验室检查": {},
      "病理检查": {},
      "影像学检查": {}
    },
    "评估": "",
    "处理": "加用射频电疗、中频脉冲理疗"
  },
  "第二次病程记录": {
    "主诉": "经过上述治疗, 患者无大碍, 准备出院",
    "体检": "血压控制稳定, 下肢疼痛缓解",
    "辅检": {
      "实验室检查": {},
      "病理检查": {},
      "影像学检查": {}
    },
    "评估": "腰椎间盘膨出保守治疗后症状缓解, 高血压2级 (很高危组) 控制稳定",
    "处理": "准许出院"
  }
},
"最终诊断": [
  "腰椎间盘膨出",
  "高血压2级 (很高危组)"
],
"最终诊断依据": "\n1、腰椎间盘膨出。病史: 患者主因左下肢放射性疼痛1月余入院, 疼痛活动时减轻、休息后加重, 并逐渐累及腰部。症状: 右下肢放射性疼痛, 以腓肠肌为著, 伸展活动受限。体征: 左侧下肢皮肤感觉较右侧敏感, 直腿抬高试验阳性。辅助检查: 腰椎MRI显示腰椎退行性病变, L3/4、L4/5椎间盘膨出, 椎管狭窄。2、高血压2级 (很高危组)。体征: 血压160/100mmHg。辅助检查: 心电图示窦性心律, 正常心电图 (排除心脏继发性病变)。"}

```

Figure 15: Chinese EHR example.

```

{"id": "pmc-6146624-1",
"basic_information": "Female, 62",
"chief_complaint": "Pain and a sense of weight in the right hypochondriac site",
"past_medical_history": "HBV-related hepatopathy",
"present_medical_history": "Patient has been suffering for a few months from pain and a sense of weight in the right hypochondriac area. A CT scan reveals a lesion of about 5 cm between the VI hepatic segment and right colon. Colonoscopy shows an ab extrinsec compression near the right colonic flexure that dislocates the bowel, complicating the endoscope's progression.",
"physical_examination": "Not provided in the text",
"auxiliary_examination": {
  "Imaging examination": {"CT scan": "Lesion of about 5 cm between the VI hepatic segment and right colon"},
  "Laboratory examination": {"AFP": "394.90 ng/ml (normal value <15)","CEA": "2.20 ng/ml (normal value <5)","HBV Antibodies": "Positive"},
  "Pathological examination": {"Laparoscopic findings": "An exophytic, encapsulated neof ormation with a diameter of 5 cm at the level of the VI hepatic segment"}},
"clinical_department": "Hepatobiliary & Pancreas Surgical Dept.",
"preliminary_diagnosis": ["Hepatic adenoma","Hepatocellular carcinoma","Colorectal cancer"],
"preliminary_diagnosis_basis": "1. Hepatic adenoma: history: The patient has HBV-related hepatopathy, and has a lesion in the liver; symptoms: Pain and a sense of weight in the right hypochondriac site; signs: An exophytic, encapsulated neof ormation found in the liver segment; auxiliary_examination: CT scan shows a 5 cm lesion, and elevated AFP levels. 2. Hepatocellular carcinoma: history: Patient's liver condition with HBV and elevated AFP levels; symptoms: Similar symptoms as hepatic adenoma; signs: abnormal findings in imaging; auxiliary_examination: markedly elevated AFP levels indicate a higher risk for liver cancer. 3. Colorectal cancer: history: The colonoscopy revealed compression and difficulty in progression; symptoms: Potential bowel obstruction signs; signs: Extrinsic compression observed during colonoscopy.",
"differential_diagnosis": "1. Cholangiocarcinoma: May present with biliary obstruction symptoms and elevated liver function tests. 2. Metastatic liver disease: Symptoms may include weight loss and jaundice, with imaging showing multiple lesions in the liver. 3. Gallbladder carcinoma: Can cause similar symptoms, but typically with gallbladder-related signs.",
"first_treatment": "Surgical resection of the hepatic lesion with close monitoring of liver function and AFP levels postoperatively.",
"course_record": {
  "first_course_record": {
    "chief_complaint": ["The patient has been suffering from pain and a sense of weight in the right hypochondriac site for a few months."],
    "physical_examination": ["Tenderness in the right upper quadrant, no rebound tenderness."],
    "auxiliary_examination": {
      "Imaging examination": {"CT scan": "A lesion of about 5 cm was found between the VI hepatic segment and right colon."},
      "Laboratory examination": {"AFP": "394.90 ng/ml (normal value <15)","CEA": "2.20 ng/ml (normal value <5)"},
      "Pathological examination": {}},
    "assessment": "Based on the complaints and examination results, there is evidence supporting the diagnosis of Hepatocellular carcinoma due to ongoing pain and tenderness in the right upper quadrant, alongside the need to consider Hepatitis B virus infection as a contributing factor.",
    "treatment": ["Patient was scheduled for laparoscopic resection of the lesion in the VI hepatic segment."]
  },
  "second_course_record": {
    "chief_complaint": ["The patient reports improvement in abdominal pain after the procedure."],
    "physical_examination": ["Abdomen soft with mild tenderness in the right upper quadrant, no significant findings."],
    "auxiliary_examination": {
      "Imaging examination": {},
      "Laboratory examination": {},
      "Pathological examination": {}},
    "assessment": "Based on the complaints and examination results, the assessment supports the ongoing diagnosis of Hepatocellular carcinoma, while also confirming the presence of Hepatitis B virus infection as a significant factor in the patient's condition.",
    "treatment": ["The patient underwent laparoscopic resection of the segment without complications."]
  },
  "third_course_record": {
    "chief_complaint": ["The patient feels well and would like to be discharged."],
    "physical_examination": ["Vital signs stable, no tenderness in the abdomen, recovery appears satisfactory."],
    "auxiliary_examination": {
      "Imaging examination": {},
      "Laboratory examination": {},
      "Pathological examination": {}},
    "assessment": "Based on the complaints and examination results, there is no evidence to suggest a change in the diagnosis of Hepatocellular carcinoma or Hepatitis B virus infection; therefore, the current assessments remain unchanged.",
    "treatment": ["The patient was permitted to leave the hospital the day after the operation."]
  }
},
"final_diagnosis": ["Hepatocellular carcinoma","Hepatitis B virus infection"],
"final_diagnosis_basis": "1. Hepatocellular carcinoma: history: Patient has HBV-related hepatopathy; symptoms: Pain and a sense of weight in the right hypochondriac site; signs: CT scan shows a 5 cm lesion in the VI hepatic segment; auxiliary_examination: Elevated AFP level of 394.90 ng/ml. || 2. Hepatitis B virus infection: history: Positive antibodies to HBV; symptoms: N/A; signs: N/A; auxiliary_examination: N/A."}

```

Figure 16: English EHR example.