

CliMedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models in Clinical Scenarios

Zetian Ouyang¹ Yishuai Qiu¹, Linlin Wang^{1*}, Gerard de Melo²,
Ya Zhang³, Yanfeng Wang³, Liang He¹

¹East China Normal University ²Hasso Plattner Institute

³Shanghai Jiao Tong University

{51265901102,51265901063}@stu.ecnu.edu.cn, {llwang,lhe}@cs.ecnu.edu.cn,
gdm@demelo.org, {ya_zhang,wangyanfeng}@sjtu.edu.cn

Abstract

With the proliferation of Large Language Models (LLMs) in diverse domains, there is a particular need for unified evaluation standards in Chinese clinical medical scenarios, where models need to be examined very thoroughly. We present CliMedBench, a comprehensive benchmark with 14 expert-guided core clinical scenarios specifically designed to assess the medical ability of LLMs across 7 pivot dimensions¹. It comprises 33,735 questions derived from real-world medical reports of top-tier tertiary hospitals and authentic examination exercises. The reliability of this benchmark has been confirmed in several ways. Subsequent experiments with existing LLMs have led to the following findings: (i) Chinese medical LLMs underperform on this benchmark, especially where medical reasoning and factual consistency are vital, underscoring the need for advances in clinical knowledge and diagnostic accuracy. (ii) Several general-domain LLMs demonstrate substantial potential in medical clinics, while the limited input capacity of many medical LLMs hinders their practical use. These findings reveal both the strengths and limitations of LLMs in clinical scenarios and offer critical insights for medical research.

1 Introduction

With the advent of Chinese medical large language models (LLMs) such as HuatuoGPT (Zhang et al., 2023), ChatMed (Zhu and Wang, 2023), and BenTsao (Wang et al., 2023), the potential for these tools in healthcare has expanded considerably (Singhal et al., 2023). These models are engineered to address intricate medical problems by providing diagnostic assistance and treatment suggestions. Nonetheless, the absence of a comprehensive and systematic evaluation of their performance, which encompasses aspects such as re-

sponse accuracy, hallucination incidence, and content safety, hampers their integration into clinical practice. Consequently, there is an urgent need for a standardized evaluation benchmark to scrutinize the capabilities and limitations of such medical LLMs.

Developing a practically relevant benchmark is non-trivial. There is a substantial disconnect between current benchmarks for the Chinese language and the realities of medical practice, as such benchmarks are mostly derived from open educational resources (Mbakwe et al., 2023). A benchmark based on real-world medical cases offers superior authenticity and heterogeneity, while more accurately mirroring the intricacies encountered in clinical practice. These cases present greater challenge and complexity, leading to a more rigorous assessment of model performance and robustness in practical applications, including clinical decision support, diagnosis, and treatment recommendations. Moreover, benchmarks developed from open resources are susceptible to data contamination issues.

Prevalent medical benchmarks like MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022) incorporate data from accessible sources such as textbooks, scholarly articles, and qualification examinations. The effectiveness of such evaluation benchmarks is controversial: medical exams are inefficient clinical performance indicators. Large-scale EHR-based benchmarks such as emr-QA (Pampari et al., 2018) have addressed the deficiency in clinical QA; however, language discrepancies preclude their direct applicability for evaluating Chinese medical LLMs. Chinese benchmarks, including CMExam (Liu et al., 2024), CMB (Wang et al., 2024), and MLEC-QA (Li et al., 2021), primarily source their data from exams such as CNMLE and NMLEC. MedBench (Cai et al., 2024) uses exam questions and artificially generated EHRs to evaluate the LLMs' exam-solving

*Corresponding author.

¹<https://github.com/Optifine-TAT/CliMedBench>

capabilities in different subdisciplines. Despite their comprehensive analysis, these benchmarks are disconnected from actual medical practice due to their lack of real-world medical case data, and the use of exam-based data raises concerns about data pollution. The need for good benchmarks makes evaluating performance a significant challenge.

To address the limitations of prior research, we introduce the CliMedBench, a robust benchmark comprising 33,735 questions across 14 core medical scenarios assessing LLMs’ ability across six dimensions, primarily sourced from authentic cases to align with medical standards and practices. The CliMedBench integrates expertise from Chinese medical practitioners, offering a valid measure for gauging medical linguistic proficiency and cognitive skills in LLMs. We evaluate various general and medical-specific LLMs using this benchmark and perform a comprehensive analysis that sheds light on relevant research avenues to enhance the medical capabilities of LLMs. The main findings on this benchmark are as follows:

- Chinese medical LLMs underperform on this benchmark, especially where medical reasoning and factual consistency are vital, underscoring the need for advances in clinical knowledge and diagnostic accuracy.
- Several general-domain LLMs demonstrate substantial potential in medical clinics, while the limited input capacity of many medical LLMs hinders their practical use.
- The indeterminacy inherent in medical contexts can significantly compromise the accuracy of model-generated responses.

2 The Proposed Benchmark

2.1 The Taxonomy of CliMedBench

A well-structured taxonomy enables us to conduct a more fine-granular assessment of medical LLMs, while also ascertaining that the evaluation is comprehensive and practically relevant. Our taxonomy, designed to maintain the benchmark’s applicability and comprehensiveness in real-world clinical scenarios, is based on a categorization that mirrors medical practice and fully covers it, as inspired by Liang et al. (2022). As depicted in Figure 1, we build on a “Who–What–How” scheme to categorize real-world clinical medical practice, providing

14 core clinical scenarios for assessment. Along the “Who” axis, we distinguish five principal roles in the medical field: the *Radiographer*, *Pharmacist*, *Patient*, *Medical Student*, and *Specialist Doctor*, where doctors encompass attending physicians, surgeons, and other medical specialists. “What” addresses a broad spectrum of key medical scenarios, covering basic knowledge tests, in-hospital diagnosis, clinical pathway reasoning, case summaries, wrong treatment detection, etc. This allows CliMedBench to evaluate the medical ability of LLMs from seven perspectives, including clinical question answering, knowledge application, reasoning, information retrieval, summarization abilities, hallucination, and toxicity.

To illustrate the divisions within CliMedBench, we provide an example of the scenario mappings for In-hospital Diagnoses (ID) in Table 1. ID is one of the core scenarios in CliMedBench that spans four periods, encapsulating the patient care continuum from admission to discharge, and consider the following scenario descriptions:

- ID #1 refers to the selection of examinations by healthcare professionals and radiographers.
- ID #2 involves the diagnosis by physicians, integrating examination results with the patient’s medical history and additional health data.
- In the ID #3 period, treatment strategies, ranging from pharmacological interventions to surgical procedures, are developed in collaboration with pharmacists and medical staff.
- ID #4 pertains to physicians providing discharge instructions to patients.

Who	What (Task)
Doctor, Patient	ID#1
Doctor, Radiographer, Patient	ID#2
Doctor, Pharmacist, Patient	ID#2
Doctor, Patient	ID#4

Evaluation axes: clinical QA ability

Table 1: Example mappings in clinic scenarios. In subsequent sections, tasks will be designated using acronyms formed by the initial letters.

2.2 Construction and Statistics

CliMedBench is derived from real-world Electronic Health Records (EHRs) of top-tier tertiary hospi-

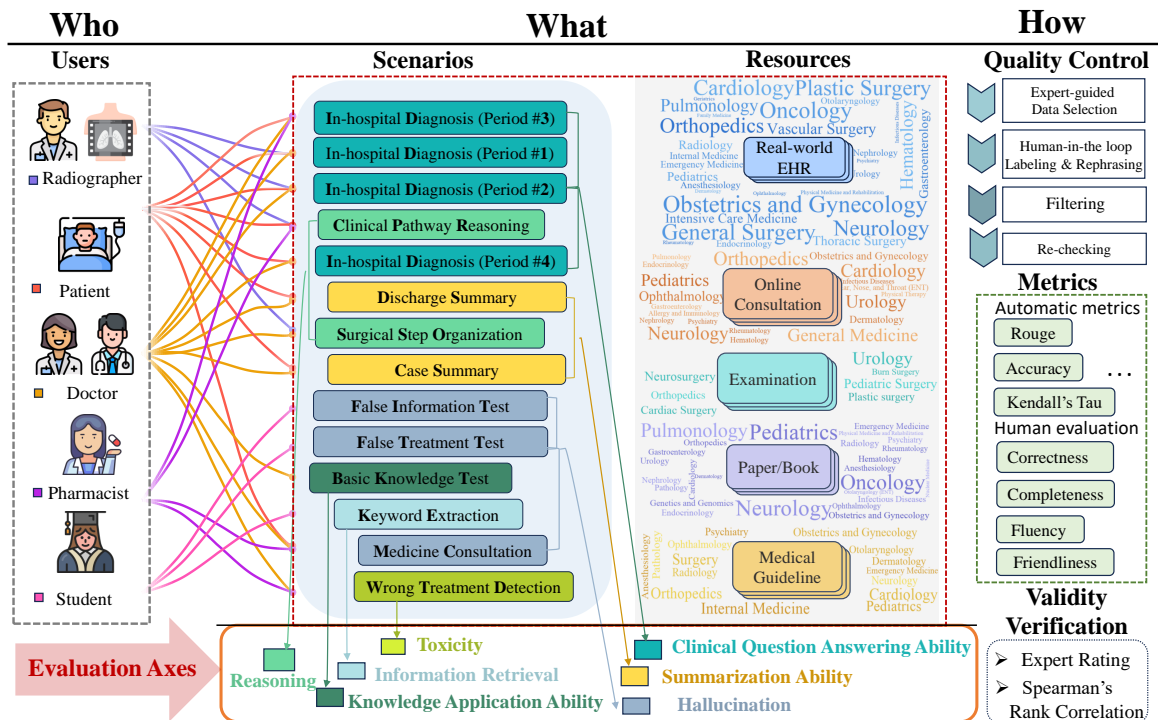


Figure 1: Overview of CliMedBench with “Who-What-How” taxonomy linking users with core clinical scenarios.

tals in China, supplemented with examination exercises, medical guidelines, textbooks, scholarly articles, and human-annotated online consultations. This corpus spans a multitude of medical specialties, meticulously curated to enhance the diversity of CliMedBench. The EHR data were supplied by our collaborating hospital. The dataset comprises all EHRs that include radiological records from the period January 1, 2023, to March 31, 2023. Examination questions are derived from the NMLEC 2023 Annual Examination, covering 16 subjects within the domains of Surgery and Internal Medicine.

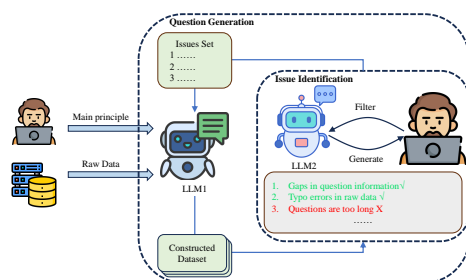


Figure 2: Workflow of collaboration between humans and LLMs for dataset construction.

Before the dataset construction, all EHRs have been doubly de-identified by ethics committees and experts to make sure no PHI of patients or health-care professionals is leaked. Then we preprocess the raw data to filter low-quality data and ensure

proper formatting by automatic techniques such as regular expression matching and medical entity recognition. Figure 2 illustrates the workflow of our construction. Medical experts initially develop guidelines based on the content types indicated by doctors’ notes in EHRs; these guidelines are used to segment EHRs into sentences, which are then categorized by LLM1 into components for question generation, with QA-pairs formulated by coupling correct answers with distractors sourced from disparate but thematically similar EHR segments. A secondary LLM (LLM2) audits the dataset for inconsistencies or ambiguities, with flagged items evaluated by medical experts to ensure only relevant issues are addressed. Feedback from this evaluation refines LLM1’s processing to mitigate recurrent dataset flaws. The dataset undergoes iterative validation cycles, ensuring a minimum of 90% of the questions meet quality standards as confirmed by dual expert review. We finally obtained 33,735 instances for 14 core clinical scenarios that are strictly based on doctor’s notes and clinical treatment recordings. Figure 3 depicts the data distribution of CliMedBench, encompassing 19 branches of medicine, e.g., neurosurgery and gastroenterology. CliMedBench has three question types, including:

- Multiple-choice clinical question answering.
- Sequencing questions, e.g., surgical step re-

ordering.

- Open-ended generation, e.g., discharge summary, subjective clinical question answering.

To confirm the effectiveness of benchmark construction, we employ diverse methodologies to validate CliMedBench, as described in Section 6.

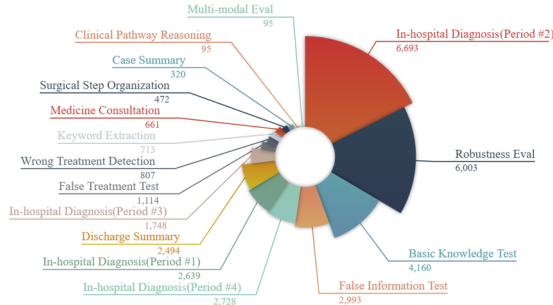


Figure 3: Data distribution of clinical scenarios.

2.3 Characteristics

CliMedBench improves over existing benchmarks in several respects: (1) **Authenticity and Uniqueness**, as it genuinely reflects doctors’ practical experience by exclusively using expert-annotated EHRs from top hospitals with up-to-date, authentic insights, while reducing the potential of data contamination. (2) **Comprehensiveness and Multi-dimensionality**, as it is meticulously designed to align with Chinese clinical practices, encompassing diverse medical disciplines with multimodal information, offering a broad spectrum of evaluation perspectives. (3) **Practicality**, as it offers a novel agent-based Computerized Adaptive Testing approach to guarantee rapid assessment with CliMedBench.

3 Models and Evaluation Metrics

To assess the state-of-the-art, we conduct evaluations using CliMedBench of 11 representative LLMs from both the general and medical domains, including OpenAI’s GPT series² (gpt-3.5-turbo-1106 and gpt-4-1106-preview), ChatGLM3 (Zeng et al., 2023), ERNIE-Bot 4.0³, Xinghuo v3⁴, Qwen max⁵, Baichuan⁶, HuaTuoGPT (Zhang et al., 2023), BenTsaio (Wang et al., 2023), MedicalGPT⁷, and ChatMed (Zhu and Wang, 2023). The

²<https://chat.openai.com>

³<https://yiyan.baidu.com>

⁴<https://xinghuo.xfyun.cn>

⁵<https://tongyi.aliyun.com>

⁶<https://github.com/baichuan-inc/Baichuan-13B>

⁷<https://github.com/shibing624/MedicalGPT>

base model of HuatuoGPT, HuatuoGPT2, Bentsao, and MedicalGPT are Ziya-LLaMA-13B-Pretrain-v1, Baichuan2-13B-Chat, Alpaca-Chinese-7B, and Baichuan-13B-Chat, respectively. We adopt the default or publisher-recommended parameter settings in their published website. Given the presence of multiple-choice, sequencing, and open-ended generation questions in CliMedBench, we utilize a comprehensive set of metrics. Specifically, we use Accuracy for multiple-choice question answering and Kendall’s τ (Kendall, 1938) for sequencing questions. For open-ended generation, we combine expert-level human evaluation with supplementary automatic evaluation metrics, e.g., ROUGE-1 (Lin, 2004) for discharge summary and SimCSE-based similarity for wrong treatment assessment. For the latter, we first apply fine-tuning to SimCSE (Gao et al., 2021) using distinct medical documents, then utilize the resulting model to derive sentence vectors, and finally compute the semantic similarity with the reference.

4 Main Results

We conduct an in-depth evaluation of 11 LLMs using CliMedBench, stringently examining their performance across seven pivot dimensions. Corresponding comparisons utilizing automatic metrics are provided in Table 2. We also engage human experts to assess open-ended generation (WTD and multi-modal report analysis) across four dimensions in Figure 4, including medical correctness, completeness, fluency, and friendliness.

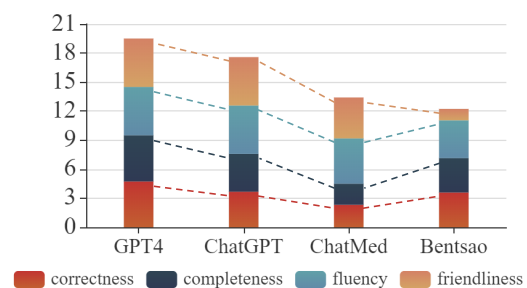


Figure 4: Human evaluation results of four aspects.

Chinese medical LLMs underperform on this benchmark, especially where medical reasoning and factual consistency are vital. Comparative analysis reveals that models via APIs generally outperform others, with average scores exceeding 50. ERNIE-Bot, GPT-4, and Qwen achieve fairly similar average scores of 69.2, 69.0, and 68.5, respectively. In contrast, current medical LLMs exhibit notably inadequate performance: Even the best-

	Model	ID#3	ID#1	ID#2	CPR	ID#4	DS	SSO
		ACC.	ACC.	ACC.	Kendall's τ	ACC.	ACC.	Kendall's τ
General	GPT4	87.8	68.4	97.4	73.2	84.6	98.2	77.0
	ChatGPT	76.8	86.3	97.4	59.5	70.6	85.4	42.6
	ERNIE-Bot	78.3	87.4	98.7	79.5	83.3	94.2	67.13
	SparkDesk	65.3	85.0	98.7	61.5	53.0	26.6	30.4
	Qwen	84.6	89.4	95.0	69.6	85.6	97.1	67.1
	Baichuan	47.6	56.7	88.5	22.1	31.2	32.1	23.9
	ChatGLM3	47.2	88.0	97.6	33.5	40.6	60.4	21.1
Specialized	HuatuogPT	26.6	48.0	66.6	24.7	25.6	20.3	3.4
	BenTsao	27.2	24.6	24.6	25.2	4.6	1.0	18.8
	MedicalGPT	41.3	43.7	81.4	39.5	31.0	20.4	21.7
	ChatMed	13.6	37.4	20.6	4.5	8.6	2.8	1.5
	Model	CS	FIT	FTT	BKT	KE	MC	WTD
		ACC.	ACC.	ACC.	ACC.	ROUGE-1	ACC.	Similarity
General	GPT4	98.4	25.0	12.6	70.8	40.2	44.0	81.3
	ChatGPT	97.1	2.8	1.3	51.9	39.8	38.9	80.9
	ERNIE-Bot	99.7	13.5	10.7	79.8	42.0	53.3	81.9
	SparkDesk	95.6	11.7	4.0	68.7	28.8	63.5	81.0
	Qwen	99.1	13.9	13.3	82.4	39.7	49.2	80.4
	Baichuan	73.4	1.7	21.2	38.8	33.6	37.1	78.6
	ChatGLM3	92.0	9.6	6.8	46.9	34.4	45.5	78.8
Specialized	HuatuogPT	61.2	13.8	8.6	22.6	29.3	23.0	79.2
	BenTsao	25.6	0.6	0	20.6	6.5	27.6	75.2
	MedicalGPT	67.2	1.9	7.1	35.0	33.3	41.7	77.1
	ChatMed	10.9	2.0	1.2	9.3	11.4	12.4	75.8

Table 2: Results of 11 LLMs with automatic metrics on the 14 core clinic scenarios of CliMedBench.

performing MedicalGPT only achieves an average score of 38.7. This deficiency primarily stems from the substandard language understanding capabilities of those LLMs.

Several Chinese LLMs (ERNIE-Bot and Qwen) demonstrate performance on par with GPT-4 in clinical medicine of China, achieving scores primarily within the range of 68.5 to 69.2. This could stem from the unique treatments, expression styles, and China-manufactured pharmaceuticals, which diverge from what is encountered in the training data of GPT series models. A disparity of capabilities between these Chinese LLMs and the GPT series predominantly manifests in medical knowledge and reasoning.

Next, we will summarize the performance of LLMs with regard to particular evaluation dimensions. Regarding clinical QA abilities, Qwen outperforms others with an average score of 88.7. However, variability in model performance across scenarios is evident, with ChatGPT achieving the highest score (97.4) on ID #2 but not ranking among the top performers in other scenarios. GPT4 and ERNIE-Bot show exceptional reasoning capabilities, achieving average scores of 75.1 and 73.3, respectively. The notable performance disparity between general and medical-specific LLMs high-

lights the need for further enhancement in the reasoning ability of medical-specific LLMs. In all evaluated models, hallucinations are significantly pronounced. The FIT data is designed to trigger hallucinations by incorporating an erroneous reference. Their data sources are the same as BKT, however, model accuracy exhibits a marked reduction, plummeting from an average score of 47.3 to 8.3. This substantial decline shows the vulnerability of LLMs to uncritically adopt perspectives presented in their input, highlighting an immediate need for enhancement. Hallucinations exhibited on the WTD dataset indicate that for questions with special structures, LLMs not only need to master the knowledge points examined by the questions but also need to understand the logical relationships in the questions, which may exceed the ability of the models. The knowledge application ability of the leading general LLMs ranges from 79.8 to 82.4, suggesting a substantial reservoir of medical knowledge of these models. For the information retrieval task, MedicalGPT has significantly narrowed the disparity with leading LLMs, achieving a score of 33.3, merely 8.7 points below the top-performing ERNIE-Bot. This improvement predominantly stems from the specialized nature of its generated terminologies.

5 Quantitative Analysis

Chain-of-Thought To demonstrate the potential improvement of LLM’s reasoning abilities with customized prompts, we compare the performance of four representative models using vanilla and Chain-of-Thought (CoT) prompts in Figure 5.

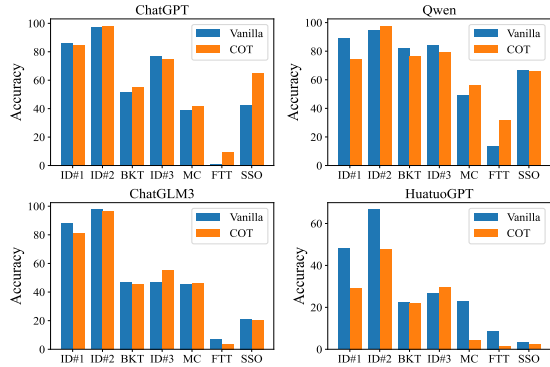


Figure 5: Accuracy comparison of four models on seven datasets using both vanilla and CoT prompts.

We observe that the utilization of tailored CoT prompts significantly enhances model performance across seven datasets that demand higher reasoning skills. Specifically, for Qwen, there is an average accuracy increase from 62.7% to 69.2%. It suggests that CoT can enhance reasoning and hallucination resistance in medical contexts, as observed in the Surgical Organization, False Info Test, and Wrong Treatment datasets. Conversely, the impact of CoT prompts on ChatGLM3 is minimal and it adversely affects huatuoGPT, underscoring the dependency of CoT prompt efficacy on the model’s comprehension and contextual correlation proficiency. In addition, the long text of the few-shot CoT prompt (on average 4.987 times longer than the vanilla prompt) is also a reason for the decrease in accuracy, as described in the following paragraph.

Limited Input Capacity We notice that EHRs frequently contain a variety of diagnostic test outcomes, records of prior treatments, and familial and social histories, often spanning multiple pages. Consequently, the limited input capacity of many LLMs poses a challenge to their practical use in clinical scenarios. Figure 6 depicts performance comparisons across varying input windows.

We observe a notable decrease, declining from 47.3 to 43.1, in the performance of nearly all LLMs as the length of the inputs increases, revealing that the limited input capacity is the main factor hindering their performance in clinical medicine. In addition, medical LLMs exhibit a more pronounced

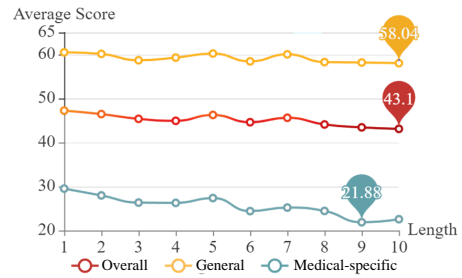


Figure 6: Performance across varying input windows, where the x-axis represents dataset segments, 1 being the shortest and 10 the longest, sorted by length.

decline (from 29.5 to 22.6) compared to the general LLMs (from 60.5 to 58.0), suggesting that these specialized LLMs may be less capable of maintaining performance with longer inputs.

Robustness Test To conduct a robustness test, we introduce manually-crafted perturbations that comprise shape-based character conversion, homophonic substitution, simplified-to-traditional Chinese transformation, and random symbol insertion. These perturbations cover 12% of the characters. Figure 7 provides the robustness test results.

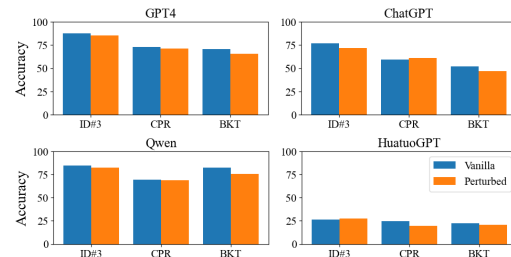


Figure 7: Robustness test of GPT4, ChatGPT, Qwen, and HuatuoGPT on different datasets.

We observe that with perturbations, all models exhibited a reduction in scores, ranging from 2.0 to 3.2, with particularly notable decreases observed on the basic knowledge test scenario, averaging at 4.7. This shows the significant impact of even minor disturbances on model performance despite their seemingly negligible impact on readability.

Multi-modal Capability To further investigate the performance of models in multi-modal settings, we have compiled a set of 92 diagnostic image pairs from medical textbooks (primarily consisting of MRI and ultrasound scans) and PathVQA (He et al., 2020) to assess the potential of LLMs in multimodal medical diagnosis. Our evaluation primarily focuses on the representative model GPT-4V. GPT-4V does not achieve satisfying results in these cases, and only 16.7% of its responses are rele-

vant to the reference. Figure 8 depicts an example case, where GPT-4V demonstrates its utility by successfully identifying inflammation in the patient’s shoulder tendon through ultrasound imaging diagnostics. However, detailed but crucial diagnostic indicators within the images remained virtually indiscernible. Moreover, GPT-4V often disclaims diagnostic capability and withholds responses. Therefore, its utility in multi-modal clinical scenarios is currently limited.

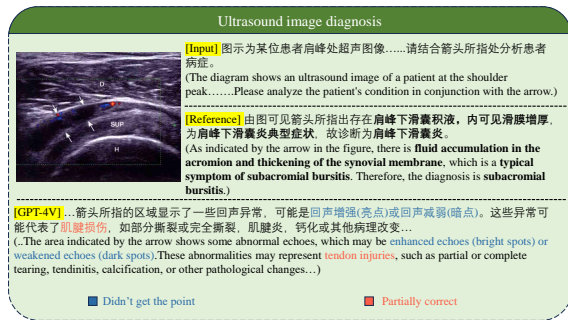


Figure 8: A multi-modal example case.

Inadequate Instruction Following Ability During the assessment, we observe that LLMs exhibit inadequate instruction following ability, a deficiency particularly conspicuous within medical-specific LLMs. Even in straightforward multiple-choice scenarios, several medical-specific LLMs, e.g., ChatMed and BenTsao, struggle to follow the given instructions to accomplish the task. Therefore, the average scores for ChatMed and Bentsao are only a quarter to a sixth of that achieved by the top-performing model. This highlights the necessity of enhancing the model’s aptitude for comprehending and following diverse instructions, thereby enabling adaptation to different tasks.

Potential Causes of Toxicity When evaluating the toxicity, we find that general LLMs, guided by safety protocols, often err on the side of caution, indiscriminately flagging and inhibiting potentially hazardous medical actions, including some that are clinically justified. Conversely, medical-specific LLMs disproportionately focus on the potential benefits of medical interventions, often neglecting potential patient-specific repercussions, e.g., advising a feverish patient who recently consumed alcohol to consider taking acetaminophen. This one-sided approach by each model type leads to suboptimal performance in toxicity assessment.

Lack of Innovative Thinking As experts pointed out in the human assessment, the responses

of LLMs on CliMedBench significantly lack innovation. To quantify this, we substitute the option of the correct answer in multiple-choice questions with “None of the above is correct” (Pal et al., 2022). Surprisingly, we find that this triggers a fall in accuracy to less than 10% in the false treatment test scenario for the majority of models. This suggests that, with instruction tuning, LLMs often opt for a seemingly reasonable choice from the given alternatives, potentially overlooking more precise solutions in clinical scenarios, thereby limiting the innovative capacity of medical LLMs.

6 Benchmark Validity Verification

To ensure the real-world applicability and validity of CliMedBench in clinical settings, the design rationale of the 14 core scenarios is meticulously aligned with actual clinical practices. For instance, the **In-hospital Diagnosis** datasets are carefully structured to adhere to established clinical pathways in healthcare settings. To further confirm the reliability of CliMedBench, we engaged medical professionals to assess our benchmark from three perspectives, including medical accuracy, assessment effectiveness, and language proficiency. Figure 9 plots the assessment results, which substantiate the quality of CliMedBench with an “acceptable” (3 points) or higher rating.

Furthermore, we calculate the Spearman correlation between our CliMedBench and another representative benchmark, namely MedBench⁸ based on other kinds of data. This approach⁹ allows us to conduct multidimensional evaluations that reflect both collective and discrete correlations between benchmarks. Figure 9 illustrates a robust correlation between the CliMedBench and MedBench leaderboards, with an overall Spearman correlation of 0.943 and subdivisions no less than 0.657, substantiating CliMedBench’s utility and reliability as an evaluative benchmark.

7 The Agent-based Computerized Adaptive Testing (CAT) Approach

During the evaluation phase, we identified two key issues: (1) Smaller LLMs struggle with exceedingly difficult questions, resulting in uniformly low accuracy and a lack of differentiation in the evaluations. (2) Certain LLMs exhibit slow GPU in-

⁸<https://medbench.opencompass.org.cn/leaderboard>

⁹https://github.com/ctlllll/understanding_llm_benchmark?s?tab=readme-ov-file

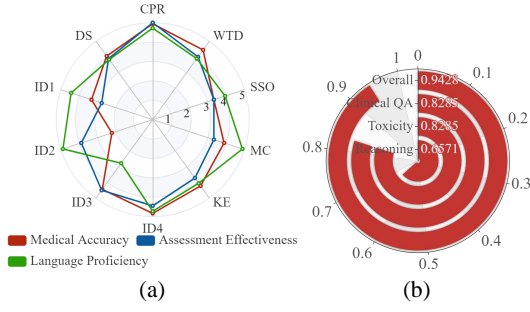


Figure 9: Figure (a) depicts the assessment results of medical experts, while Figure (b) shows a noteworthy correlation between CliMedBench and MedBench.

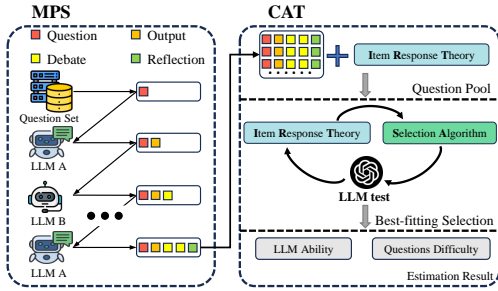


Figure 10: The workflow of Agent-based CAT.

ference speeds or high API-related computational costs, significantly increasing benchmark testing expenses. To address this issue, we propose an agent-based CAT approach, enabling rapid assessment of model performance using CliMedBench.

Theoretical Basis Our approach is fundamentally rooted in the psychometric Item Response Theory (IRT). We incorporate the three-parameter logistic model (IRT-3PL), formulated as:

$$P(X_{ij} = 1|\theta_j) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \quad (1)$$

where a_i , b_i , and c_i represent the discrimination, difficulty, and the guessing factor, respectively. θ_j represents the proficiency of LLM j , and $P(X_{ij} = 1|\theta_j)$ is the probability that an LLM j with proficiency θ_j gives a correct response to question i .

Procedure of Agent-based CAT As depicted in Figure 10, our agent-based CAT consists of two main steps: Multi-Agent Based Participant Synthesis (MPS) and Computerized Adaptive Testing (CAT). MPS leverages multi-agent LLMs to synthesize data that mimics participant behavior, which is used to construct a question pool. We then use CAT to sequentially select the best-fitting questions from the question pool to evaluate the ability of LLMs.

Algorithm 1 provides the “generation-debate-reflection” process in MPS for data synthesis, aiming to overcome the difficulty of insufficient participant data in previous CAT (Zhuang et al., 2023). Once a sufficient number of participant behavior data is synthesized, the modeling process aligns more closely with the assumptions of IRT. Specifically, we use permutations of 5 LLMs (e.g., Bloomz and ChatGLM2) to form multi-agent based participants, serving as the examinees in the IRT process to synthesize performance-related data.

Algorithm 1 The MPS process

- 1: **Input:** Q =Question data;
- 2: **Output:** Synthesize data
- 3: **function** MULTI-AGENT(Q, LLM_1, LLM_2)
- 4: **for** Q_i in Q **do**
- 5: $G \leftarrow LLM_1$ -Generation(Q_i)
- 6: $D \leftarrow LLM_2$ -Debate(Q_i, G)
- 7: $R \leftarrow LLM_1$ -Reflection(Q_i, G, D)
- 8: DebateResults $\leftarrow (Q_i, G, D, R)$
- 9: **end for**
- 10: **return** Synthesized data
- 11: **end function**

To accomplish the best-fitting selection, the CAT step includes two components that work alternately, including (1) ability estimation using IRT in Eq. 1 and (2) question selection via Fisher information.

Method	Accuracy
CAT (Zhuang et al., 2023)	32.10
Our Agent-based CAT	40.74

Table 3: Comparison to other CAT method.

Results We select 243 questions from CliMedBench to conduct a rapid assessment using our agent-based Computerized Adaptive Testing (CAT). To validate the effectiveness of our agent-based CAT rapid assessment, we compare its results with the regular CliMedBench evaluation, which involved 33,735 instances as proposed in Section 2. This comparison is illustrated in Figure 11. Our observations indicate a consistency in the relative rankings of LLMs derived from the two evaluation methods, validating the effectiveness of using a limited set of questions to gauge model ability.

Table 3 compares the accuracy of our agent-based CAT with that of the representative previous work by Zhuang et al. (2023). We observe a relative performance increase of 26.9%, demonstrating the efficacy of MPS in synthesizing sufficient data

Benchmark	Language	Data Source	Coverage	Size
CliMedBench	Chinese	EHR(major), NMLEC, and Books	14 Core Scenarios	33.7k
CMExam	Chinese	CNMLE	Exam QA	60k
CMB	Chinese	Exams and Chinese Medical Question Database	Exam QA	281k
MLEC-QA	Chinese	NMLEC and Books	Exam QA	136k
RJUA-QA	Chinese	EHR	Urology	2.1k
MedBench	Chinese	CNMLE and NMLEC	Exam QA	40k
NLPEC	Chinese	NLPEC	Exam QA	21.7k
MultiMedQA	English	Existing datasets	-	18.8k
PubMedQA	English	PubMed	Biomedical paper QA	1k labeled
emrQA	English	EHR	EHR QA	400k

Table 4: Contrasts between the CliMedBench and existing benchmarks.

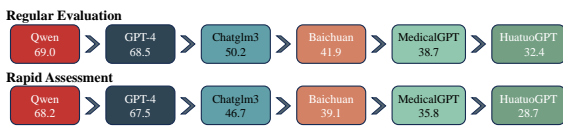


Figure 11: Comparisons between regular CliMedBench evaluation and our agent-based CAT rapid assessment.

that better aligns with the IRT assumption.

8 Related Work

MultiMedQA (Singhal et al., 2023) and PubMedQA (Jin et al., 2019) are effective benchmarks based on QA tasks for evaluating the medical abilities of LLMs. Large-scale EHR-based benchmarks such as emr-QA (Pampari et al., 2018) have addressed the deficiency in clinical QA, but language discrepancies preclude their direct applicability for evaluating Chinese medical LLMs. The assessment of LLMs for Chinese medical proficiency has traditionally relied on benchmarks derived from multiple-choice and generative question-answering formats, utilizing resources like exam questions, textbooks, and doctor-patient interactions. Chinese benchmarks, including CMExam (Liu et al., 2024), CMB (Wang et al., 2024), MLEC-QA (Li et al., 2021), and MedQA (Jin et al., 2021), primarily source their data from exams such as CNMLE and NMLEC. Despite their comprehensive analysis, these benchmarks are disconnected from actual medical practice due to their lack of real-world medical case data, and the challenge of ensuring quality control and avoiding data pollution grows proportionally with the involved volume of data (Li et al., 2023). Li et al. (2020) collected more than 21k multi-choice questions from the National Licensed Pharmacist Examination in China, but only the test set of the dataset has been released.

Finally, we contrast our work with other lines of work sharing seemingly similar goals. MedBench (Cai et al., 2024) is an exhaustive benchmark designed for the domain of Chinese medical QA, it utilizes exam questions and synthetical EHRs to evaluate the LLMs’ exam-solving capabilities in different areas, rather than actual clinical skills. In contrast, our benchmark extends this framework across 14 diverse medical scenarios. Furthermore, despite MedBench providing preliminary empirical analysis, it lacks in-depth qualitative analyses of the model’s performance. RJUA-QA (Lyu et al., 2023) creates high-quality medical datasets to evaluate clinical reasoning based on EHRs and clinical cases. However, it is restricted to urology, offering limited insight into the broader medical capabilities of LLMs. Table 4 delineates the contrasts between the CliMedBench and existing benchmarks from the perspectives of language, data sources, coverage, and size.

9 Conclusion and Discussions

This paper introduces CliMedBench, a robust benchmark derived from real medical cases that comprises 33,735 questions across 14 core medical scenarios assessing LLMs’ ability across six dimensions. Evaluating diverse LLMs reveals their suboptimal performance, especially where medical reasoning and factual consistency are vital, underscoring the need for advances in clinical knowledge and diagnostic accuracy. We also conducted a comprehensive qualitative analysis of the experimental outcomes and made several novel insights. Simultaneously, we proposed the agent-based CAT approach, which enables rapid assessment with minimal problem sets.

10 Limitations and Ethical Issues

Protected Health Information (PHI) encompasses data related to an individual’s health status, health-care provision, or payment for healthcare services, which is generated or amassed by a Covered Entity or its Business Associate. PHI typically undergoes de-identification to safeguard individual privacy prior to the dataset’s publication. CliMedBench is a dataset derived primarily from real-world medical cases and the Chinese National Physician Qualification Examination. All EHRs and codes have been doubly de-identified by ethics committees and experts according to the guidance and have passed the ethical review of our partner hospitals before submission. However, such real-world data may suffer from noise. This stems from two main sources: (i) erroneous data input by medical personnel during recording or formatting error during data retrieval, and (ii) inaccuracies introduced in automatic information extraction. Users should exercise caution regarding data reliability in light of these limitations. In future work, extensive validation by medical experts will be conducted to ensure the correctness of all data. Our project has been conducted in collaboration with relevant medical centers with proper approval of all data sharing. Due to legal restrictions, our data is available for research purposes only. Researchers can contact us with the research objectives and intended use of the data. We ensure full compliance with applicable laws and ethical guidelines during data collection and use, all information in medical cases has been desensitized to ensure that no personal information related to patients or medical personnel is leaked.

References

- Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2024. MedBench: A large-scale Chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17709–17717.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- 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 Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Maurice Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.
- Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, and Anqi Wang. 2020. [Towards medical machine reading comprehension with structural knowledge and plain text](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1427–1438, Online. Association for Computational Linguistics.
- Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023. Huatuo-26m, a large-scale Chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.
- Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A chinese multi-choice biomedical question answering dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yan Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksesgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. [Holistic evaluation of language models](#). *ArXiv*, abs/2211.09110.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu

- You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.
- Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, et al. 2023. RJUA-QA: A comprehensive QA dataset for urology. *arXiv preprint arXiv:2312.09785*.
- Amarachi B Mbakwe, Ismini Lourentzou, Leo Anthony Celi, Oren J Mechanic, and Alon Dagan. 2023. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS digital health*, 2(2):e0000205.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikandan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.
- Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. [HuaTuo: Tuning LLaMA model with Chinese medical knowledge](#). *arXiv 2304.06975*.
- Xidong Wang, Guiming Chen, Song Dingjie, Zhang Zhiyi, Zhihong Chen, Qingying Xiao, Junying Chen, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. [CMB: A comprehensive medical benchmark in Chinese](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6184–6205, Mexico City, Mexico. Association for Computational Linguistics.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2023. GLM-130B: An open bilingual pre-trained model. *ICLR 2023*.
- Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. [HuatuoGPT, towards taming language model to be a doctor](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.
- Wei Zhu and Xiaoling Wang. 2023. ChatMed: A Chinese medical large language model. <https://github.com/michael-wzhu/ChatMed>.
- Yan Zhuang, Qi Liu, Yuting Ning, Weizhe Huang, Rui Lv, Zhenya Huang, Guan Hao Zhao, Zheng Zhang, Qingyang Mao, Shijin Wang, et al. 2023. Efficiently measuring the cognitive ability of llms: An adaptive testing perspective. *arXiv preprint arXiv:2306.10512*.