MedBench: A Large-Scale Chinese Benchmark for Evaluating Medical Large Language Models

Yan Cai\textsuperscript{1}, Linlin Wang\textsuperscript{1,2,}\textsuperscript{*}, Ye Wang\textsuperscript{3}, Gerard de Melo\textsuperscript{3,4}, Ya Zhang\textsuperscript{2,5}, Yanfeng Wang\textsuperscript{2,5}, Liang He\textsuperscript{1}

\textsuperscript{1}East China Normal University
\textsuperscript{2}Shanghai Artificial Intelligence Laboratory
\textsuperscript{3}Hasso Plattner Institute
\textsuperscript{4}University of Potsdam
\textsuperscript{5}Shanghai Jiao Tong University

\{51255901104,yewang\}@stu.ecnu.edu.cn, \{llwang,lhe\}@cs.ecnu.edu.cn, gdm@demelo.org, \{ya_zhang,wangyanfeng\}@sjtu.edu.cn

Abstract

The emergence of various medical large language models (LLMs) in the medical domain has highlighted the need for unified evaluation standards, as manual evaluation of LLMs proves to be time-consuming and labor-intensive. To address this issue, we introduce MedBench, a comprehensive benchmark for the Chinese medical domain, comprising 40,041 questions sourced from authentic examination exercises and medical reports of diverse branches of medicine. In particular, this benchmark is composed of four key components: the Chinese Medical Licensing Examination, the Resident Standardization Training Examination, the Doctor In-Charge Qualification Examination, and real-world clinic cases encompassing examinations, diagnoses, and treatments. MedBench replicates the educational progression and clinical practice experiences of doctors in Mainland China, thereby establishing itself as a credible benchmark for assessing the mastery of knowledge and reasoning abilities in medical language learning models. We perform extensive experiments and conduct an in-depth analysis from diverse perspectives, which culminate in the following findings: (1) Chinese medical LLMs underperform on this benchmark, highlighting the need for significant advances in clinical knowledge and diagnostic precision. (2) Several general-domain LLMs surprisingly possess considerable medical knowledge. These findings elucidate both the capabilities and limitations of LLMs within the context of MedBench, with the ultimate goal of aiding the medical research community.

Introduction

The advent of large language models (LLMs) has demonstrated substantial potential for diverse real-world applications, thanks to their remarkable language understanding capabilities. In the medical domain, a notable number of Chinese medical LLMs have successively emerged, including HuaTuo (Wang et al. 2023), BianQue (Chen et al. 2023), and DoctorGLM (Xiong et al. 2023), to better assist doctors in diverse tasks ranging from clinical diagnosis to disease prevention (Singhal et al. 2023). This underscores an urgent need for a standardized medical benchmark, capable of offering reliable and authoritative evaluations for such LLMs. Assessing the potential and inherent limitations of medical LLMs from diverse perspectives continues to present considerable challenges (Singhal et al. 2023; Chang et al. 2023).

The primary cause of this issue lies in the pronounced discrepancy between the existing benchmarks and the practical realities of medicine, leading to an urgent need for advances in evaluation standards. Widely used medical benchmarks such as MedQA (Jin et al. 2021) and MedMCQA (Pal, Uma-pathi, and Sankarasubbu 2022) typically encompass publicly available medical question-answering datasets gathered from textbooks, research papers, and board exams. However, in light of recent research employing these open-access datasets for model training (Han et al. 2023), a process potentially resulting in data contamination with regard to the evaluation (Nori et al. 2023), it is increasingly clear that the prevailing benchmarks exhibit considerable limitations. A recent study has made advances by introducing a human evaluation framework along with MultiMedQA (Singhal et al. 2023), a multi-faceted benchmark that combines existing QA datasets with a novel set of online medical questions. Nonetheless, given the disparities among typologically diverse languages, we are unable to rely on this benchmark for assessing Chinese large language models, in particular considering the unique clinical standards and procedures across different countries illustrated in Figure 1.

Hence, several Chinese benchmarks have been proposed,
including a medical NLP-task oriented one\(^1\) and others based on the Chinese National Medical Licensing Examination (CNMLE), including MLEC-QA (Li, Zhong, and Chen 2021) and CMExam (Liu et al. 2023). While these benchmarks provide valuable insights, they are not exhaustive and may fall short in enabling us to comprehensively gauge the potential of LLMs with respect to all-round medical knowledge and their practical utility in real-world clinical diagnostic scenarios. In mainland China, a unique three-stage examination process is employed, key components of which, including the Resident Standardization Training and Doctor in-charge Qualification Exams, have largely been overlooked in prior work. Furthermore, the evaluation of real-world clinical practical skills has not been adequately incorporated in prior work.

To address these gaps and align with the learning and growth trajectory of Chinese doctors, we introduce MedBench, a novel large-scale Chinese medical benchmark, which encompasses both authentic three-stage medical examinations and real-world clinical diagnosis cases. It surpasses prior benchmarks by being exclusively sourced from the latest validated exams and expert-annotated EHRs, thereby ensuring compliance with medical standards and practices. Furthermore, we conduct extensive experiments and offer detailed analyses to provide diverse perspectives for evaluating clinical knowledge recall and reasoning capabilities of LLMs across a range of branches of medicine. The main findings on this benchmark are as follows:

- Chinese medical large language models underperform on this benchmark, necessitating substantial improvements in clinical knowledge and diagnostic accuracy, as well as refinement of their original in-context learning abilities.
- Several large language models designed for general-domain tasks possess substantial medical knowledge, thereby exhibiting promising potential.
- Human evaluation reveals that ChatGPT possesses rich knowledge for clinical practices, while current Chinese medical LLMs lack high-quality conversational abilities and sufficient medical knowledge.
- ChatGPT’s unsatisfactory performance in case analysis, a question type that demands profound medical knowledge understanding and clinical reasoning abilities, reveals significant room for improvement in medicine.

In summary, this study presents a comprehensive benchmark that aligns with the practical realities of medicine in mainland China and provides profound empirical findings to reveal the medical capabilities and limitations of LLMs. Furthermore, we incorporate Item Response Theory (Birnbaum 1969) to further enhance our benchmark, with the ultimate goal of aiding the medical research community.

### The Proposed Benchmark

**Overview**

For better evaluating medical foundation models, this paper proposes MedBench as a large benchmark with 40,041 exercises originating from both authentic medical examinations and real-world diagnostic and treatment cases. Specifically, we collect three-stage medical examinations that serve as a faithful reflection of the comprehensive process involved in obtaining medical licenses in mainland China, exemplifying essential medical knowledge. Moreover, we construct a number of real-world cases based on electronic health records that provide examination plans, diagnoses, and treatments based on patients’ symptoms, which can reveal the medical knowledge utilization and reasoning capabilities of LLMs in the real world.

**Construction and Statistics**

Regarding the three-stage medical examinations, we collect representative exercises from the Chinese Medical Licensing Exam (CNMLE), Resident Standardization Training Exam, and Doctor in-charge Qualification Exam\(^2\) across many recent years, covering 27,248, 2,841, and 8,927 questions, respectively. The examination exercises, depicted on the left side of Figure 2, are

\(^1\)https://github.com/michael-wzhu/PromptCBLUE

\(^2\)http://www.nmec.org.cn/
multiple-choice questions that can be categorized into three types:

- A1/A2/B: Single statement questions with one correct answer out of five options.
- A3/A4: A series of questions accompanied by a clinical case with one correct answer out of five options.
- Case Analysis: Given a clinical case, a series of questions are created with 6–12 options per question. Some questions may have more than one correct answer.

Figure 3 depicts the classification of branches of medicine for some examinations in MedBench. Note that we use internal medicine and surgery as representative examples for the latter two stages, considering the vast number of subcategories within these fields.

Furthermore, we collect over 2,000 real-world electronic health records, and employ experts to identify the symptoms, diagnoses, treatments, and examinations from the reports, resulting in 701 high-quality ones with an average length of 60–100 words per report (Figure 2). The annotators completed professional training and strictly following predefined annotation standards to ensure accuracy. Based on these annotated reports, a total of 1,025 question–answer pairs are eventually formed.

Characteristics MedBench surpasses existing benchmarks in several aspects: (1) Authenticity and Novelty. It exclusively leverages expert-annotated EHRs and authentic up-to-date medical examinations to mitigate contamination. (2) Comprehensiveness and Multi-facetedness. It is meticulously designed to align with Chinese medical standards and practices by incorporating three-stage multi-disciplinary examinations and real-world clinical cases. (3) Practicality. Human evaluation on clinical real-world cases ensures congruence with practical realities of medicine, while difficulty-stratified divisions in MedBench enable rapid assessment.

Evaluation

Models and Evaluation Metrics To evaluate the medical capabilities, we conduct assessments using MedBench with several representative LLMs from both the general and medical domains, including ChatGPT, ChatGLM (Zeng et al. 2023; Du et al. 2022), Baichuan-13B, HuaTuo, and ChatMed. Furthermore, we evaluate other Chinese medical LLMs, such as BianQue, but we have observed that it lacks the ability to deliver accurate and reasonable responses to multiple-choice questions.

For the three-stage multiple-choice examinations, we employ accuracy as the evaluation metric. When dealing with real-world cases, we combine expert-level human evaluation with the additional automatic evaluation metrics BLEU (Papineni et al. 2002) and ROUGE (Lin 2004).

Experimental Settings We conduct extensive experiments to evaluate the five-shot performance of LLMs, ensuring their capability to respond in a multiple-choice format. We leverage the API for ChatGPT and opt for local deployment to facilitate evaluations for other LLMs. Furthermore, we partition the MedBench dataset based on exams, medical subdiscipline, and question types, and perform independent testing on each subset to enable a comprehensive analysis.

Main Results

Three-stage examination results. In Table 1, we present a comprehensive analysis of the accuracy metrics for various LLMs across the three exams. A salient observation is that ChatGPT consistently surpasses other models, despite the latter being training on extensive Chinese corpora or premium medical datasets. Nevertheless, our observations indicate that ChatGPT’s accuracy rate hovers around 50% for CNMLE and approximately 60% for other assessments, exposing substantial avenues for enhancement.

As depicted in Figure 5, ChatGPT exhibits subpar performance on questions pertaining to Traditional Chinese Medicine (TCM) and Chinese Western Medicine (CWM), with accuracy metrics oscillating between 40-45%. Conversely, LLMs trained on a more expansive Chinese dataset demonstrate a narrower performance disparity on TCM and CWM questions relative to ChatGPT. This suggests that a contributing factor to ChatGPT’s diminished efficacy on MedBench may be attributed to its limited exposure to Chinese data during its pretraining phase, consequently affecting its proficiency in Chinese medical knowledge.

It is imperative to underscore that Baichuan-13B and HuaTuo encounter difficulties in case analysis questions, highlighting potential domains for refinement in their logical reasoning or multi-turn dialogue competencies.

We also perform comparative evaluations with a number of recently-developed models, including MedLLaMA (Wu et al. 2023; Du et al. 2022), Baichuan-13B, HuaTuo, and ChatMed. Furthermore, we evaluate other Chinese medical LLMs, such as BianQue, but we have observed that it lacks the ability to deliver accurate and reasonable responses to multiple-choice questions.
The objective of this evaluation endeavor was to quantitatively assess the correctness, completeness, fluency, and friendliness of HuaTuo, ChatMed, ChatGPT, and GPT-4. Figure 4 provides the outcomes of this human evaluation. GPT-4 is found to consistently manifest superior performance across all delineated criteria, while ChatGPT marginally trails GPT-4, particularly on correctness and completeness. Conversely, HuaTuo’s rating in friendliness is somewhat diminished, predominantly attributable to sporadic generation of incongruous content. However, it is important to note that HuaTuo demonstrates pronounced levels of correctness and completeness, evincing profound medical knowledge, though its articulation warrants improvement. ChatMed, while laudable for its fluency and friendliness, registers suboptimal scores in correctness and completeness. Such disparities might be indicative of a potential attenuation of medical expertise during the fine-tuning phase.

Quantitative Analysis

Comparison on different question types. Figure 5 provides the detailed results on CNMLE, the Resident Standardization Training Exam (denoted as “Resident”) and the Doctor in-charge Qualification Exam (“In-charge”). Clearly, ChatGPT continues to exhibit a substantial performance advantage over the other LLMs. A3/A4 questions and case analysis questions, characterized by their multi-question structure, present a substantial challenge to the conversational abilities of the models. If a given LLM fares poorly on these question types, it indicates its inherent limitations in efficiently managing multi-turn dialogue or questions necessitating intricate reasoning steps.

Chain-of-Thought In the experimental investigation, it is observed that prompts tailored to the LLM have the potential to enhance the reasoning capabilities of the LLM. Figure 7 illustrates the choice distribution of Baichuan-13B under both the vanilla prompt and Chain-of-Thought prompt-
Figure 5: Results for three-stage medical examinations.
为什么选择这个选项 (Why choose this option) ?

因为血/骨髓涂片是疟疾确诊的“金标准”，可以检测出患者血液或骨髓中的疟原虫来确诊疟疾。其他选项如放射免疫测定、病史、血常规和腹部B超也可以作为疟疾确诊的参考指标，但它们不能像血/骨髓涂片那样直接检测疟原虫。所以，选项E是正确的。

回答 (Answer): E. 疟疾 (Malaria)

为什么选择这个选项 (Why choose this option) ?

因为根据病史和体征，患者可能患有伤寒。伤寒是一种由伤寒杆菌引起的传染病，常发生在饮食不卫生的人群中。根据病史和体征，患者可能有持续高热、乏力、腹泻、肌肉和关节疼痛、结膜充血等症状。血常规检查可见白细胞计数升高，分类可见异型淋巴细胞，支持伤寒的诊断。因此，选项A是最合适的选择。

回答 (Answer): A. 伤寒 (Typhoid fever)

Fewer questions, higher differentiation. During the assessment process, some shortcomings became apparent: (1) Certain LLMs manifest suboptimal inference speed on GPUs or high computational cost when interfacing via an API. Given that MedBench comprises approximately 40,000 questions, performing inference on the entirety of this dataset consecutively takes considerable time for LLMs. (2) While categorization of questions based on the type provides a rudimentary gauge of difficulty, it is important to acknowledge that questions within the same category can exhibit disparate levels of difficulty. (3) Assigning particularly challenging questions to less capable LLMs is inadvisable, as it may culminate in uniformly diminished accuracy, making meaningful distinctions impossible.

To ameliorate these challenges, we propose methodological strategies to classify questions of analogous types according to their inherent difficulty gradients. By adopting this paradigm, we can optimize the evaluation process, allowing LLMs to undertake inference on a curated set of questions. As a result, this allows for a more nuanced alignment between LLMs and questions of commensurate difficulty tiers, congruent with their individual proficiencies. Specifically, we introduce an advanced evaluative framework for LLMs drawing inspiration from Item Response Theory (IRT). Our approach integrates the three-parameter logistic model (IRT-3PL), given by:

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

Here, \( \theta_j \) represents the proficiency of LLM \( j \), and \( P(X_{ij} = 1 | \theta_j) \) is the probability that an LLM \( j \) with proficiency \( \theta_j \) gives a correct response to question \( i \). This equation highlights three crucial parameters for each question \( i \): discrimination \( (a_i) \), difficulty \( (b_i) \), and the guessing factor \( (c_i) \). In our approach, we assume constant values for both \( a_i \) and \( c_i \), and then group items based on the shared difficulty metric \( b_i \). For a specific difficulty subset, these three parameters remain unchanged. The equation captures the LLM’s probability \( \theta_j \), determined from the inference results when interacting with a specific subset.

A total of 7,335 questions were divided into 10 difficulty levels. To validate question differentiation, we analyze them using BLOOMZ-7.1B (Muennighoff et al. 2022), Qwen-7B\(^6\), ChatGLM-turbo (Zeng et al. 2023), Qwen-max (Bai et al. 2023) across various difficulty levels. Part (a) of Figure 8 displays the accuracy trends of these LLMs across different prompts, where (a) and (b) depict the results using vanilla and Chain-of-Thought prompt, respectively.
these levels, highlighting a decrease in accuracy with increasing difficulty. Figure 8(b) shows the difference in accuracy between BLOOMZ-7.1B and Qwen-7B, as well as between Qwen-max and ChatGLM-turbo, for different levels of question difficulty. It is noticeable that the differences between the LLMs are significant when the question difficulty is appropriate.

Reasoning Abilities

In MedBench, there exist several typical types of reasoning:

- **Multi-condition single-hop reasoning**: A type of reasoning that requires the LLM to engage in single-hop reasoning based on a clinical scenario.

- **Statement identification**: A type of reasoning that requires the LLM to judge the correctness of multiple statements, which further tests the LLM’s knowledge base.

- **Multi-hop reasoning**: A type of reasoning with multiple questions, where the questions are related, requiring the LLM to perform multi-hop reasoning.

To ascertain the extent to which LLMs make accurate decisions based on their grasp of pertinent knowledge, we extended our inquiry, which entailed requesting the LLMs to furnish justifications for their responses, as depicted in Figure 6. Our observations reveal that LLMs can substantiate their answers when they are accurate. Conversely, in cases where erroneous responses are given, the accompanying explanations often prove illogical, which provides compelling evidence that these LLMs either lack the requisite knowledge in the domain or are incapable of rationalizing towards the correct solutions.

Related Work

Traditionally, medical LLMs relied on classic medical QA benchmarks for evaluation. Some studies used USMLE for evaluation and achieved satisfactory results (Kung et al. 2023; Nori et al. 2023), with zero-shot GPT-4 achieving average scores of 86.65% and 86.7% on the Self-Assessment and Sample Exam parts of USMLE, respectively. However, questions from USMLE have a very distinct Western medical perspective that is notably different from that of Chinese medicine. As two distinct medical systems, the latter includes traditional Chinese medicine, with obvious differences in concepts, diagnosis, and treatment methods. Therefore, the same symptoms may lead to different diagnoses and treatments. As a result, the USMLE cannot serve as an adequate benchmark for Chinese medical practice. MedQA, PubMedQA (Jin et al. 2019; Liévin, Hother, and Winther 2022), and MLEC-QA are also useful benchmarks for medical QA, comprising a large number of high-quality questions, but they were proposed early on, potentially leading to data contamination. Although MedMCQA is a relatively new benchmark, its questions originate primarily from Indian medical institutions. As a result, analogous to USMLE, it lacks Chinese medical content and thus cannot serve to evaluate Chinese medical question answering sufficiently well. CMExam is a novel benchmark for medical QA in Chinese, comprising the latest exam questions from CNMLE, which has been manually annotated by medical experts with assistance from ChatGPT to ensure high quality. However, CMExam only comprises questions from CNMLE, while overlooking other major medical exams in mainland China. Moreover, as a predominantly multiple choice dataset with only a small portion of fill-in-the-blank questions, CMExam lacks real-world medical data and scenarios. This constrains its ability to fully assess LLMs’ medical question answering capabilities. In fact, the optimal evaluation approach is to evaluate LLMs manually (Xu et al. 2023b). A recent work (Tu et al. 2024) employs specialist physicians to assess the quality. However, manual evaluation is a time-consuming and labor-intensive task, making it difficult to conduct on a large scale.

Conclusion and Discussions

In this paper, we present MedBench, an exhaustive benchmark specifically designed for the domain of Chinese medical question answering. Preliminary empirical analyses underscore the suboptimal performance of Chinese medical LLMs when subjected to this benchmark, highlighting the need for improved clinical acumen and diagnostic precision. Furthermore, the adeptness of these LLMs in contextual learning requires further refinement. During our empirical investigations, we found that certain models manifest pronounced hallucinatory behavior. As our research progresses, it is imperative to ensure data veracity while delving into the systematic evaluation of such hallucinatory phenomena. Furthermore, the appraisal of the LLM’s inferential competencies, as presented in this research, points to the need for further methodological refinements. Within the realm of clinical diagnostics, a diagnosis is typically predicated upon a plethora of corroborative evidence. In subsequent work, we plan to compile an enriched dataset, encompassing patients’ antecedent medical records and comprehensive physical examination narratives, to strengthen the evaluative framework for medical LLMs. Simultaneously, our findings underscore the efficacy of psychometric methodologies in aiding the evaluation. Moving forward, we intend to enhance and further integrate such methodologies into our model assessment paradigm.
Acknowledgments
This work was supported by the National Innovation 2030 Major S&T Project of China (No.2022ZD0160700 & 2022ZD0160705) and the National Natural Science Foundation of China under Grant 12105105. The computations were performed in the ECNU Multifunctional Platform for Innovation (001). We also thank the financial support from Shanghai Chinafortune Co., Ltd.

References


