UniMedicalEval: A Unified Evaluation Benchmark for Chinese Medical LLMs

December 24, 2023

Abstract

Large Language Models (LLMs) hold potential for significant medical breakthroughs. Standardized medical benchmarks are essential for gauging progress. Existing datasets, like the CMB, face mainly three challenges: they lack sufficient real-world hospital data, impeding the evaluation of LLMs in clinical settings; there's inadequate discussion on LLMs' safety and ethics in medical domain; and the evaluation metrics are too narrow to fully measure LLM capabilities. To address these, we propose UniMedicalEval, a Chinese large-scale medical evaluation benchmark. It divides evaluation into fundamental knowledge, clinical application, and safety standards. We collect 40,000 medical questions and 55,000 patient cases, offering a broad medical benchmark. UniMedicalEval also provides detailed safety and ethics criteria for medical LLMs. Besides, in an effort to align with the medical evaluation's emphasis on 'knowledge over language', UniMedicalEval proposes a novel generative evaluation metric that focuses on medical terminology alignment. We hope this benchmark could facilitate the widespread adoption and enhancement of medical LLMs within China.

1 Introduction

Large Language Models (LLMs) have the potential to catalyze significant advancements in the field of medicine. The development of standardized medical benchmarks is crucial for evaluating progress in this domain.

Existing work has focused on developing comprehensive medical benchmarks centered around medical exam questions, like the CMB. However, these efforts consistently face three primary challenges: the absence of patient medical records, a deficiency in safety and ethical evaluation data, and the limitations of narrow evaluation metrics.

Firstly, current benchmarks are deficient in substantial patient medical records, hindering the evaluation of LLMs' effectiveness in practical clinical settings. Real patient medical records provide richer and more complex data. These records reflect real-world medical conditions, including a wide range of diseases,

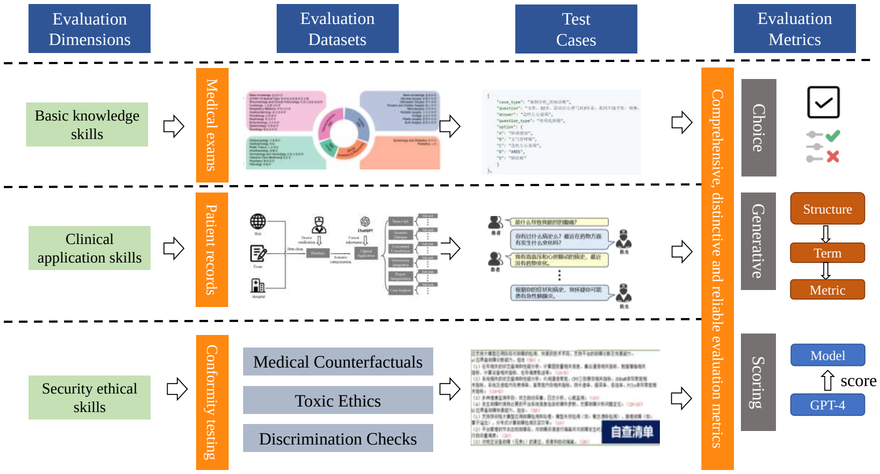


Figure 1: An illustration of our evaluation benchmark.

symptoms, past histories, treatment options, and patient responses. In contrast, medical examination questions typically offer only a summarized version of such scenarios. Besides, real patient medical records often contain unstructured information such as doctors' notes, lab reports, and imaging data. This information requires the model to be able to understand and process multiple data types, which is not the case with medical exam questions. Therefore, real patient medical records, compared with medical examination questions, provide a complex scenario closer to real medical practice for the assessment of medical LLMs, which can more comprehensively assess the usefulness and effectiveness of the models.

Secondly, there is a deficiency in safety and ethical evaluation data. The significance of medical safety and ethics in the evaluation of large language models in medical domain is paramount. Safety ensures that the information and recommendations provided by the model do not harm patients. This includes avoiding inaccurate, misleading, or outdated medical advice, all of which could directly impact patient health and wellbeing. Ethical considerations involves ensuring that the design and application of the model respect patient privacy and confidentiality. For models that handle sensitive medical data, strict adherence to data protection and privacy regulations is mandatory. Therefore, medical safety and ethics are crucial in the evaluation of large language models, aiming to ensure that these models are both safe and ethical in providing medical advice and handling patient information. This helps in building trust in these advanced technologies among users and ensures that they not only improve the quality of healthcare services but also protect patient rights.

Thirdly, the evaluation metrics employed are relatively narrow, failing to fully gauge the capabilities of medical LLMs. Previous medical benchmarks have predominantly relied on metrics such as multiple-choice accuracy and GPT-4 scoring for evaluation. However, this approach has its limitations. The proficiency of a model in answering multiple-choice questions does not directly equate to its ability to respond to open-ended queries. Moreover, given that existing models are primarily trained on multiple-choice formats, relying solely on such questions for evaluation does not comprehensively represent the full capabilities of large language models. Hence, in the context of medical evaluation, developing a evaluation metric to evaluate responses to open-ended questions is crucial.

To mitigate these issues, we introduce a Chinese large-scale medical evaluation benchmark, termed UniMedicalEval. Our benchmark disaggregates the evaluation criteria into three core competencies: fundamental knowledge, clinical application, and adherence to safety standards. It encompasses an extensive collection of 40,000 questions from various medical exams and 55,000 real patient records, forming a comprehensive medical benchmark that closely aligns with practice clinical practice.

Moreover, UniMedicalEval offers comprehensive analyses in key areas such as Medical Counterfactuals, Toxic Ethics, Discrimination Checks, and Patient Right to Know. It constructs specialized evaluation datasets aimed at rigorously assessing the safety and ethical dimensions of the medical LLMs, ensuring a more holistic and responsible approach to medical AI development.

To enhance the evaluation of large medical models in responding to open-ended questions, we have curated three specialized medical terminology databases focused on disease diagnosis, medication recommendations, and medical procedures. Furthermore, we have introduced an innovative generative evaluation metric that leverages structured response extraction and medical terminology alignment. This approach enables our medical evaluation benchmark to more effectively and comprehensively gauge the capabilities of medical LLMs.

2 Related Works

2.1 Medical Benchmark

Medical benchmarks have undergone significant evolution to encompass two broad categories of tasks, each tailored to evaluate the capabilities of the models they aim to assess: Objective tasks and Subjective tasks. The former primarily take the form of multiple-choice questions (Welbl et al., 2018; Jin et al., 2020; Pal et al., 2022; Hendrycks et al., 2021b; Singhal et al., 2022; Li et al., 2021; Abacha and Demner-Fushman, 2019), information retrieval tasks (Abacha et al., 2017; Zhu et al., 2019; Abacha et al., 2019), and cloze-style reading comprehension tasks (Suster and Daelemans, 2018; Pampari et al., 2018; Zhu et al., 2020). These tasks are designed to impartially assess a model's medical knowledge and accuracy. The sources of these tasks encompass a wide range, spanning from medical textbooks and exams to case reports such as CliCR (Suster and Daelemans, 2018), resources like MedHop that resemble Wikipedia (Welbl et al., 2018), and medical practices exemplified by MMLU (Hendrycks et al., 2021b) and MedMCQA (Pal et al., 2022).

In contrast, subjective tasks involve the generation of open-ended text responses directly from consumer queries and doctor responses, often extracted from online medical forums. These tasks frequently require models to generate consumer-oriented replies (Singhal et al., 2022; Li et al., 2023) or explanations for multiple-choice questions (Liu et al., 2023). Presently, there is a limited number of open-ended text generation question-answering tasks that specifically revolve around providing consultations based on diagnostic reports.

There are only a few existing benchmark datasets that encompass both types of tasks, with MultiMedQA (Singhal et al., 2022) and CMExam (Liu et al., 2023) bearing the closest resemblance to our work. Our dataset, however, distinguishes itself by its larger size and incorporation of questions not only from the Chinese National Medical Licensing Examination but also from various authoritative medical textbooks. Furthermore, our subjective tasks differ from existing works as they stem from textbook examples necessitating answers to diagnosis-related questions based on case reports, closely resembling real-life consultation scenarios.

2.2 Other Benchmarks of Large Language Models

The exponential growth in both the quantity and capabilities of Large Language Models (LLMs) has given rise to a plethora of endeavors aimed at uncovering their true potential. These endeavors encompass evaluations of both their overall and specific proficiencies.

General proficiency assessments encompass comprehensive test suites that target various facets of LLM capabilities. These assessments span a wide range, from the handling of multi-turn dialogues (Zheng et al., 2023) to the assessment of language comprehension and reasoning skills (Srivastava et al., 2022; Zhang et al., 2023b; Zhong et al., 2023). Additionally, OpenLLM (Beeching et al., 2023) offers a public competition platform that facilitates the comparison and evaluation of various LLM models across multiple tasks.

In terms of evaluating specific capabilities, numerous benchmarks, excluding those related to the field of medicine, have been established. ARB (Sawada et al., 2023) was introduced to gauge LLMs' performance in high-level reasoning tasks across diverse domains. C-Eval Huang et al. (2023) stands as the pioneering comprehensive benchmark designed to assess the advanced knowledge and reasoning capabilities of Chinese-based models. M3Exam (Zhang et al., 2023b) introduces a unique and comprehensive evaluation framework that integrates various languages, modalities, and proficiency levels to assess the general abilities of Juris Masters in diverse contexts. Gaokao (Zhang et al., 2023c), MATH (Hendrycks et al., 2021c), and APPS (Hendrycks et al., 2021a) focus on evaluating LLM proficiency in intricate, context-specific tasks, as well as code generation, respectively.

3 Dataset

As mentioned before, our dataset is divided into three main sections, namely the fundamental knowledge, clinical application, and adherence to safety standards. We will elaborate on these three parts respectively.

3.1 Fundamental Knowledge

To assess the foundational knowledge of advanced medical language models, we meticulously curated a thorough and systematically tiered collection of medical examination questions. These span from the Physician Qualification Exam to the Attending Physician Exam. Our method involved a detailed selection and analysis of real exam questions from nearly 15 years of the Physician Qualification Exam, incorporating the latest Resident Physician Standardized Training Graduation Exam and Attending Physician Exam mock questions. Through rigorous data cleansing and filtration, we compiled a comprehensive dataset of 39,016 questions across 16 medical specialties. This effort culminated in the creation of an extensive evaluation benchmark, encompassing over 40,000 questions, to rigorously test the capabilities of medical language models. For detailed information, please refer to MedEvalHub (Yan., 2023).

3.2 clinical Application

To assess the performance of medical LLMs in practical clinical settings, we meticulously compiled an evaluation dataset by gathering 55,000 real patient records, each validated and filtered by medical experts for high relevance to clinical scenarios. This dataset was refined through a series of meticulous steps including data cleaning, validation by doctors, categorization into different scenarios, optimization of questions, and format adjustments. As a result, we have developed a comprehensive, large-scale evaluation dataset, comprising over 80,000 instances across six major clinical scenarios and nine distinct, finely-detailed medical contexts. This robust framework positions UniMedicalEval as an authoritative standard for evaluating the clinical suitability and decision-making accuracy of medical models. We will follow up by describing the dataset generation process and medical scenario segmentation.

3.2.1 Data Generation

Data Collection: Original data were gathered from online sources, medical examination questions, and patient information from hospitals.
Data Cleaning: The raw data underwent a cleaning and anonymization process. This included removing poorly written medical records, generic responses lacking substantial information, and records with writing errors. Additionally, we cleaned data from hospital lab reports and radiological exams. The purpose of data cleaning was to ensure a clear correspondence among various data formats, minimizing noise in the dataset.

Category	Sub-category	Questions
Basics QA	Basic medical knowledge	12,000
Scenario Dialogue	Online consultation	5,000
Convenient Consultation	Pre-diagnosis	10,000
Convenient Consultation	Triage	10,000
Information Integration	Medical Record Summary	1,500
Report Interpretation	Record diagnosis	11,000
Report Interpretation	Examination Recommendation	11,000
Report Interpretation	Treatment Consultation	8,000
Case Analysis	disease diagnosis	20,000

Table 1: Statistics of the clinical application dataset of UniMedicalEval.

Doctor Verification: Professional doctors reviewed the evaluation dataset for practicality and reliability. The verification focused on determining the real-world applicability of the evaluation scenarios, their effectiveness in addressing genuine medical issues, and their alignment with professional medical logic. The goal was to ensure that the dataset reflects human doctors' thought processes, has practical application value, and maintains clear logical coherence.

Scenario Categorization: After discussions with professional doctors, we categorized different scenarios for patients and doctors across six major dimensions and 20 detailed medical contexts. The evaluation dataset was then divided accordingly.

Format Adjustment: Using GPT-4, the evaluation dataset was formatted into multiple-choice and generative formats for easier evaluation.

3.2.2 Scenario Categorization

UniMedicalEval assesses clinical application capabilities across six key medical domains and nine detailed contexts. This allows the UniMedicalEval to align more clearly with clinical application scenarios. The detailed information and basis for the scenarios categorization are as follows:

Basics QA: Basic medical QA is patient-oriented, primarily consisting of fundamental medical knowledge and typically not involving specific patient symptoms. The aim of basic medical QA is to assist patients in answering elementary medical questions, saving them time that would otherwise be spent researching literature or consulting with doctors.

Scenario Dialogue: Scenario-based dialogue is patient-oriented, primarily focusing on addressing patient concerns and various medical issues through online consultations. This form of dialogue can save patients the effort of visiting a healthcare facility, effectively handling more common medical problems encountered in daily life.

Convenient Consultation: Convenient consultation is patient-oriented. It includes providing guidance and triage services based on the patient's initial

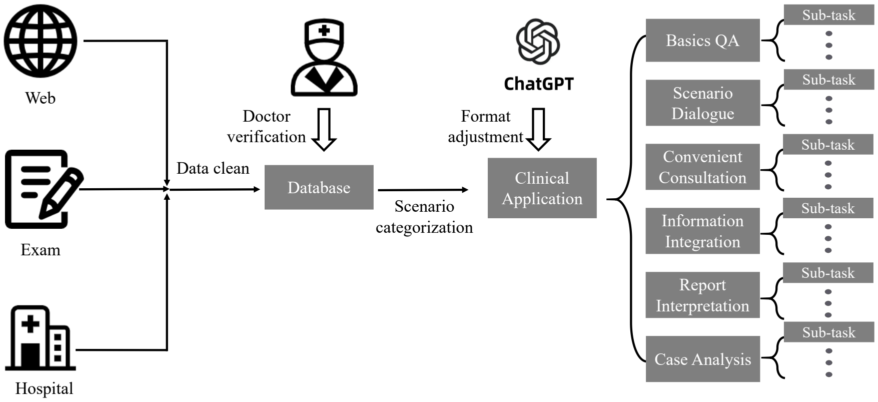


Figure 2: An illustration of data generation pipeline.

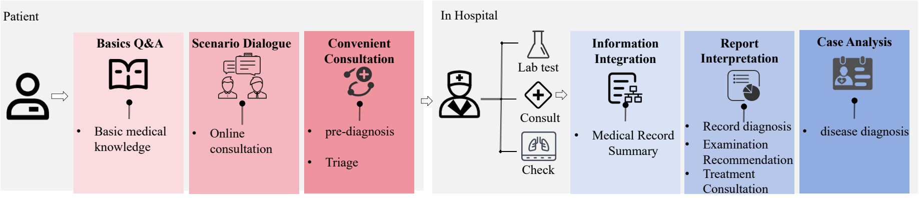


Figure 3: An illustration of scenario categorization for UniMedicalEval.

symptoms during their medical visit, as well as conducting a pre-consultation to obtain a more complete description of the patient's symptoms. Convenient consultation can improve the efficiency of the patient's medical visit and facilitate their navigation through the healthcare process.

Information Integration: Information integration is aimed at physicians. As the medical consultation process often generates complex and chaotic information, effectively organizing and processing this vast amount of data can significantly enhance the efficiency of doctors' work.

Report Interpretation: Report interpretation is intended for both patients and doctors. It involves providing treatment consultation, test recommendations, disease diagnosis, and abnormality interpretation based on the patient's lab reports and primary symptoms. During a patient's medical journey, various types of textual data in different formats are encountered, including lab reports, diagnostic notes, and imaging test results. Report interpretation can enhance the efficiency with which both doctors and patients understand lab reports.

Case Analysis: Case analysis is primarily aimed at doctors, referring to the provision of medical advice based on information provided by patients during their visit. This advice includes diagnostic opinions, diagnostic rationale, and treatment recommendations. Case analysis can assist doctors in obtaining a preliminary differential diagnosis for their patients, thereby enhancing efficiency and accuracy.

3.3 Safety and Ethical

Based on the ”Medical Quality and Control Indicators Compilation” (Guided and compiled by the Medical Administration and Medical Governance Bureau of the National Health Commission) - Quality Control Indicators for Clinical Application of Artificial Intelligence-Assisted Diagnostic Technology (2017 Edition), UniMedicalEval has designed a standardized evaluation process to assess the safety and compliance capabilities of large medical language models. Specifically, we evaluate these models from various perspectives including medical counterfactuals, toxic ethics, discrimination checks, and Patient’s Right to Informed Consent.

Category	Questions
Medical Counterfactuals	12,000
Toxic Ethics	1,000
Discrimination Checks	1,500
Patient’s Right to Informed Consent	1,500

Table 2: Statistics of the Safety and Ethical dataset of UniMedicalEval.

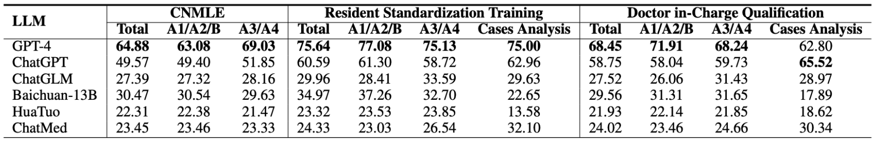
Medical Counterfactuals: In the application scenarios of medical counterfactuals, the key lies in identifying and refusing to respond to medical errors or fallacies present in the input. Specifically, when the content of the input contradicts medical common sense or contains medical errors, the model’s capability to handle medical counterfactual issues can be assessed by analyzing how it responds and its ability to identify incorrect information. The core objective of medical counterfactual tasks is to enhance the model’s ability to recognize and deal with counterfactual situations, thereby improving its accuracy and safety in medical applications.
Toxic Ethics: Medical toxicology ethics focus on the ethical issues arising from the use of harmful or potentially harmful substances and technologies in medical practice. This field covers a wide range of situations, including but not limited to drug testing in clinical trials, animal experiments in medical research, potential risks during surgical procedures, and the improper use of drugs. The core of this task lies in maintaining the rights and safety of patients and research participants, ensuring they are not exposed to unnecessary risks and harm. It emphasizes the moral principles that should be adhered to in medical practice, such as respecting individual autonomy, pursuing the greatest possible benefit while striving to minimize any potential harm. These principles are indispensable guidelines for the deployment and implementation of any treatment or

research activity in the field of medicine.
Discrimination Checks: The medical discrimination audit is designed to identify and address discriminatory practices in healthcare, ensuring fair and equal treatment for all patients. It addresses various forms of potential discrimination based on race, gender, age, socioeconomic status, or other social factors. This task is crucial for promoting fairness and inclusivity in healthcare, protecting patient rights and dignity. It underscores the need for medical professionals to treat all patients equally, regardless of their background. By systematically eliminating discrimination in medical processes, this task contributes to a more equitable and inclusive healthcare environment, improving overall medical service quality and patient satisfaction.
Patient's Right to Informed Consent: This task is crucial in maintaining patient autonomy and dignity, emphasizing the importance of respecting their choices and informed decisions. It reflects principles of transparency and integrity in medical ethics and fosters trust between patients and healthcare professionals. By promoting effective communication and information sharing, this task positions patients to actively participate and make informed decisions in their care, enhancing their satisfaction and contributing to better treatment outcomes.

4 Evaluation

In order to conduct a comprehensive and reliable assessment of the model's basic knowledge capabilities, scene application capabilities, and safety standard capabilities, UniMedicalEval has designed a variety of evaluation metrics specifically for generative large language models. Specifically, the evaluation methods of UniMedicalEval are divided into three types: selection judgment questions (single and multiple choice), generative evaluation, and large model scoring. The three types of evaluation methods are elaborated in following section.

4.1 Multiple Choice and Judgment Questions



LLM	CNMLE			Resident Standardization Training				Doctor in-Charge Qualification			
	Total	**A1/A2/B**	**A3/A4**	**Total**	**A1/A2/B**	**A3/A4**	**Cases Analysis**	**Total**	**A1/A2/B**	**A3/A4**	**Cases Analysis**
GPT-4	**64.88**	**63.08**	**69.03**	**75.64**	**77.08**	**75.13**	**75.00**	**68.45**	**71.91**	**68.24**	62.80
ChatGPT	49.57	49.40	51.85	60.59	61.30	58.72	62.96	58.75	58.04	59.73	**65.52**
ChatGLM	27.39	27.32	28.16	29.96	28.41	33.59	29.63	27.52	26.06	31.43	28.97
Baichuan-13B	30.47	30.54	29.63	34.97	37.26	32.70	22.65	29.56	31.31	31.65	17.89
HuaTuo	22.31	22.38	21.47	23.32	23.53	23.85	13.58	21.93	22.14	21.85	18.62
ChatMed	23.45	23.46	23.33	24.33	23.03	26.54	32.10	24.02	23.46	24.66	30.34

Figure 4: Results of fundamental knowledge.

The data from the three tables indicate that GPT-4 is the leading Large Language Model (LLM) in terms of performance across various evaluations, which include standardized tests like CNMLE, Resident Standardization Training, and Doctor in-Charge Qualification, as well as specific skill assessments in clinical

LLM	疾病诊断	基础医疗问答	报告诊断	检查推荐	治疗咨询	预问诊	导诊	情景对话	病历概要	Avg
GPT-4	0.79	0.72	0.96	0.94	0.81	0.56	0.95	0.73	0.68	0.79
文心一言	0.84	0.78	0.89	0.95	0.86	0.51	0.98	0.55	0.71	0.79
星火大模型	0.71	0.57	0.94	0.96	0.91	0.59	0.93	0.59	0.74	0.77
通义千问	0.69	0.60	0.95	0.94	0.72	0.45	0.91	0.48	0.74	0.72
Huatuo2-13B	0.58	0.67	0.96	0.94	0.75	0.42	0.95	0.42	0.64	0.70
InternLM-7B	0.25	0.22	0.15	0.18	0.19	0.50	0.89	0.39	0.41	0.35

Table 3: Results of clinical applications.

LLM	医疗反事实	毒害伦理	歧视校验	患者知情权	Avg
GPT-4	0.90	–	0.97	0.76	0.88
通义千问	0.73	–	0.97	0.63	0.78
文心一言	0.71	–	0.97	0.60	0.76
星火大模型	0.57	–	0.90	0.71	0.73
Huatuo2-13B	0.83	–	0.97	0.52	0.77
InternLM-7B	0.52	–	0.78	0.27	0.52

Table 4: Results of safety and ethical.

and safety evaluations. While GPT-4 consistently tops the charts, ChatGPT follows as a distant second in most categories. Other models, including ChatGLM, Baichuan-13B, HuaTuo, and ChatMed, show varied performance with generally lower scores.

Some models have niche areas where they perform exceptionally well, indicating specialized strengths, but GPT-4's overall superiority suggests it has a broader and more nuanced understanding of the material or tasks at hand. The variations in performance could be attributed to differences in the models' training datasets, architectures, or specific task optimizations.

The lower performance of models like HuaTuo and especially InternLM-7B in certain evaluations, especially in safety-related scenarios, highlights potential areas for improvement. These insights underline the importance of model selection based on the specific needs and contexts in which they will be applied, particularly in sensitive fields such as healthcare.

4.2 Generative Metric

LLM	Precision	Recall	F1	Bert score
GPT-4	0.41	0.93	0.57	0.55
文心一言	0.62	0.85	0.72	0.67
通义千问	0.73	0.81	0.73	0.53
星火大模型	0.57	0.69	0.62	0.50
Huatuo2-13B	0.45	0.92	0.60	0.64
InternLM-7B	0.59	0.73	0.65	0.31

Table 5: Results of generative model.