CMB: A Comprehensive Medical Benchmark in Chinese

Anonymous NAACL submission

Abstract

Large Language Models (LLMs) provide a possibility to make a great breakthrough in medicine. The establishment of a standardized medical benchmark becomes a fundamental cornerstone to measure progression. However, medical environments in different regions have their local characteristics, e.g., the ubiquity and significance of traditional Chinese medicine within China. Therefore, merely translating English-based medical evaluation may result in *contextual incongruities* to a local region. To solve the issue, we propose a localized medical benchmark called CMB, a Comprehensive Medical Benchmark in Chinese, designed and rooted entirely within the native Chinese linguistic and cultural framework. While traditional Chinese medicine is integral to this evaluation, it does not constitute its entirety.

Using this benchmark, we have evaluated several prominent large-scale LLMs, including ChatGPT, GPT-4, dedicated Chinese LLMs, and LLMs specialized in the medical domain. We hope this benchmark provide first-hand experience models in existing LLMs for medicine and also facilitate the widespread adoption and enhancement of medical LLMs within China.

1 Introduction

Over the past two centuries, medical advancements have substantially increased human life expectancy. Medicine's effectiveness often hinges on experience, with veteran physicians typically outperforming novices. In parallel, large language models like ChatGPT are shaped by their vast data experiences. This mutual reliance on experiential learning between physicians and LLMs suggests a promising frontier for LLMs in the medical domain.

Medical evaluation is highly professional. Although the future of *LLMs for medicine* is promising, their evaluation is a challenging topic. Deploying LLMs in hospitals raises significant ethical concerns that real-world feedback becomes difficult. Existing works on LLMs tend to leverage subjective evaluation (Zheng et al., 2023) where none of references is used during the assessment. However, the evaluation in medicine is much more professional than that of the general domain. For instance, assessing *radiology*-related issues poses a challenge for the public, a senior professor in medicine, or even a *general practitioner*. Subjective evaluation would be difficult to be scaled up since professional manual judging is expensive.

Benchmark for medical knowledge. Another school of evaluation protocol is objective evaluation, where the expected output has a clear reference. Certain protocols emphasize natural language understanding tasks that are not knowledge-intensive, as seen in studies (Zhang et al., 2022; Peng et al., 2019). In the era of Large Language Models (LLM), modern NLP evaluations underscore the significance of knowledge (Huang et al., 2023; Hendrycks et al., 2021b). In biomedicine, a typical example to probe knowledge is BioLAMA (Sung et al., 2021); however, it is tailored to evaluate masked language models instead of auto-regressive ones. Another benchmark is MultiMedBench (Tu et al., 2023), covering question answer, report summarization, visual question answering, report generation, and medical image classification. Note that MultiMedBench is only in English.

The necessity to localize medical benchmark. During economic globalization, a unified medical standard may overlook the unique medical needs and practices of different regions and ethnic groups, indicating the necessity to localize medical benchmarks. For example, in Asia, Traditional Chinese Medicine (TCM) not only offers profound insights and localized medical solutions in the prevention, treatment, and rehabilitation of diseases but also has formed a medical paradigm closely associated with regional, climatic, dietary, and lifestyle characteristics, over its long historical evolution. Simultaneously, it poses significant challenges when applying the Western medical framework to a local environment, which needs cross-cultural communication and understanding. Therefore, we should adopt a *native* medical benchmark instead of a *translated* medical benchmark for a local environment. Note that the precise translation of medical terminologies necessitates both medical professions and the cultural context in the target language.

CMB's Philosophy. The CMB dataset comprises two parts: **CMB-Exam**, featuring multiple-choice questions from qualification exams, and **CMB-Clin**, including complex clinical diagnostic questions derived from real case studies. The dataset spans 6 major categories and 28 subcategories, totaling 280,839 multiple-choice questions. For **CMB-Exam**, we selected 400 questions from each subcategory to create an evaluation set. Additionally, **CMB-Clin** is formed from 74 expert-curated medical record consultations, drawn from clinical diagnostic teaching materials. Each multiple-choice question in the dataset offers four to six options, with one or more correct answers. The clinical diagnostic questions are based on real, intricate cases, with correct answers determined by a consensus of teaching experts.

Take-away messages from CMB. After benchmarking various LLMs in CMB, we get the following observations that might be insightful. **I)** GPT-4 and recent open-sourced LLMs such as Qwen-72B-Chat and Yi-34B-Chat, have achieved an accuracy rate exceeding 60%, surpassing the threshold required for obtaining license; **II)** Accuracy exhibits significant disparities across professional levels and knowledge areas, notably between **traditional Chinese medicine** and Western medicine; **III)** The effectiveness of the **CoT and few-shot prompts** varies among models with different accuracy levels, especially presenting potential risks in knowledge-intensive tasks; and **IV)** Results of automatic evaluation using GPT-4 highly agree with **expert evaluation** results.

2 The Philosophy of CMB

2.1 The Overall Philosophy

We surveyed different medical professionals (physicians, nurses, technicians, and pharmacists) about the exams they encountered in their career development. Our research focused on common assessment types, leading us to select two key tasks for further study: multiple-choice questions and iterative questioning based on complex medical records. The former evaluates the model's knowledge grasp, while the latter assesses its practical problem-solving skills. Both tasks, having standard answers, provide reliable and stable performance indicators.

2.2 Philosophy of CMB-Exam

Existing medical benchmarks, sourced from the internet (Li et al., 2023), hospitals, etc., face privacy and accuracy challenges. We opted for qualification exams as our data source, creating the **CMB-Exam** subset. This choice is due to two key advantages: (I) qualification exams offer objective and typically accurate ground truths; (II) they provide a clear benchmark, namely a 60% accuracy rate, which corresponds to the expertise level in specific domains. The multiple-choice questions in **CMB-Exam** encompass four clinical medical professions: *physicians*, *nurses*, *medical technicians*, and *pharmacists*. These exams span the entire professional journey, from undergraduate basics, graduate selections, standardized tests, professional qualifications, to intermediate and advanced professional title exams.

In the Chinese medical field, significant work has been done on multiple-choice tasks. MLEC-QA (Li et al., 2021) compiled 21,700 manually annotated questions from the Chinese National Licensed Pharmacist Examination. Similarly, CMExam (Liu et al., 2023) gathered 68,119 tagged questions from the same exam. However, the scope of LLMs in aiding medical professions extends beyond pharmacists to include nurses, technicians, etc. Given that the Licensed Pharmacist Examination represents only a fraction of the career growth spectrum, its limited knowledge scope and occupational coverage do not provide detailed feedback. To address this, we compiled **CMB-Exam**, encompassing all medical-related occupations and the full range of exams encountered throughout their professional development.

2.3 Philosophy of CMB-Clin

Besides the theoretical exam content in **CMB-Exam**, the second subset, **CMB-Clin**, focuses on practical skills. This subset comprises complex clinical diagnostic problems to test the model's synthesis of knowledge and reasoning. It requires the model to utilize its medical knowledge for answering questions and to analyze case reports for informed responses. **CMB-Exam** and **CMB-Clin** together offer a comprehensive evaluation framework, applicable to both the career development of

Category	Subcategory	# Subject	# Questions
Physician (医师)	Resident Physician (住院医师); Licensed Assistant Physician (执业助理医师); Licensed Physician (执业医师); Associate Professional Physician (中级职称); Advanced Professional Physicians (高级职称)	81	124,926
Nurse (护理)	Practicing Nurse (护士); Licensed Practical Nurse (护师); Charge Nurse (主管护师); Advanced Practice Nurse (高级护师)	8	16,919
Technicians (医技)	Medical Technician (医技士); Medical Technologist (医技师); Supervising Technologist (主管技师)	21	27,004
Pharmacist (药师)	Licensed Pharmacist (执业西药师); Licensed TCM Pharmacist (执业中药师); Junior Pharmacist (初级药师); Junior Pharmacist Assistant (初级药士); Junior TCM Pharmacist (初级中药师); Junior TCM Pharmacist Assistant (初级中药士); Chief Pharmacists (主管药师); Chief TCM Pharmacists (主管中药师)	8	33,354
Undergraduate Disciplines (学科考试)[1]	Fundamental Medicine (基础医学); Clinical Medicine (临床医学); Traditional Chinese (TCM) and Chinese Herbal Medicine (中医学与中药学); Preventive Medicine and Public Health (预防医学与公共卫生学)	53	62,271
Graduate Entrance Exam (考研)	Integrated Western Medicine (西医综合); Integrated TCM (中医综合); Political Science (政治); Nursing (护理学)	5	16,365
Total	28	176	280,839

[1] We referenced the National Standard Subject Classification of the People's Republic of China, see https://xkb.pku.edu.cn/docs/2018-10/20220328083301969071.pdf.

Table 1: Statistics of the CMB-Exam Categories, Subcategories, Subjects, and Questions.

Split	#subcategory	#Q per subcategory	#Q in total
Test	28	400	11,200
Dev	28	10 [1]	280
Train	28	-[2]	269,359

[1] It is with explanations in dev set.
[2] Each subcategory has a different number of questions.

Table 2: Data split in CMB-Exam.

medical professionals and the learning trajectory of medical LLMs. To our knowledge, **CMB-Clin** is the inaugural multi-round question-answering dataset based on real, complex medical records.

3 Dataset Creation

3.1 Taxonomy of CMB-Exam

To obtain a precise taxonomy of medical evaluation, we aligned it with the disciplinary and examination systems of the medical field. First, we chose four main medical professions: physicians, pharmacists, medical technicians, and nurses, covering various occupational difficulty levels of examinations. Considering the learning trajectories and professional growth paths, we additionally include *discipline examinations* and *graduate entrance examinations* for these four professions, ultimately resulting in six categories: Physician, Nurse, Technician, Pharmacist, Undergraduate Disciplines, and Graduate Entrance Exam. One could refer to Table 1 for the detailed taxonomy. Moreover, we carried out a more detailed subject division within each subcategory, resulting in a total of 174 categories, the detailed directory list of which can be found in Appendix B. Through this structured arrangement, our directory structure reflects characteristics closely connected to the actual medical field, providing a solid foundation for further analysis and research.

3.2 Creation of CMB-Exam

Data Sources The data is derived from publicly available mock examination questions, coursework exercises, and summaries of commonly misunderstood examination questions. A significant portion of these materials comes from the Chinese Medical Question Database[1], from which we obtained explicit permission to share the data.

Manual Verification The data has various formats, with PDF and JSON being the most prevalent. For PDF documents, we first used Optical Character Recognition (OCR) to transform them into plain text. This text was then processed into structured formats and underwent manual verification to ensure both OCR accuracy and proper formatting.

Data Preprocessing All questions underwent a standardized data preprocessing procedure, including de-duplication and cleansing. In instances where we were unable to verify the question quality from the source, we conducted manual validation to ensure the absence of grammatical errors. Additionally, with the aid of the comment system provided by the Chinese Medical Question Database, we enacted a rigorous selection and deletion process for the data, ensuring the accuracy of the knowledge embedded in the questions.

Data Statistics Finally, we obtained a total of 280,839 multiple-choice questions. To assess the model's comprehension of medical knowledge, we randomly selected 400 questions from each subcategory as a test set. Additionally, to facilitate experiments with few-shot learning strategies, we randomly selected 10 questions from each subcategory as a development set. We then enlisted the

[1] https://www.medtiku.com/

help of three medical specialists to generate explanations for each of these questions, specifically for the purpose of conducting chain-of-thought experiments (example shown in Figure 4). The remaining 269,359 questions were used as the train set.

3.3 Creation of CMB-Clin

CMB-Clin is designed to investigate models' proficiency in knowledge application amidst real-life diagnosis and treatment circumstances.

Data Preprocessing In order to obtain a high-quality dataset, we initially collected 108 cases with questions and answers from a variety of official medical textbooks. These problems covered a wide range of disease types. Subsequently, we performed quality screening based on the following criteria: Firstly, we eliminated problems that required reliance on image information, such as those that needed CT, MRI, and other imaging data for resolution. Secondly, we selected problems that contain sufficient diagnostic information to answer the questions. Lastly, we removed cases with similar diseases to ensure the diversity of the problems. After such screening, we finally obtained 74 high-quality, complex, and real cases with 208 questions, as exemplified in Figure 1, for the construction of the CMB-Clin subset.

Task Formulation We transform the question answering task into the multi-turn dialogue task. Specifically, for each case with questions, we simulate dialogue interactions between an **examiner** and a **candidate**, focusing on assessing the model's diagnostic and therapeutic capabilities.

Figure 1 illustrates the structure of each case, which includes three main parts: **I) Description** D: patient information, including medical history summaries and chief complaints, physical examinations (e.g., visual and tactile inspection), and ancillary examinations (e.g., biopsy and CT scans); **II) Questions** Q: questions related to diagnosis and treatment based on descriptions. Some questions might be interrelated; and **III) References** R: corresponding reference answers to questions.

Formally, to simulate dialogue interactions, we concatenate the patient's description with the previous question-answer pairs and the current question in each conversation round (e.g., the k-th round). This concatenated input, denoted as x, is represented as follows: $x = D_i + Q_i + R_i + \ldots Q_{i+k}$. The reference answer for this input is R_{i+k}. For the response $\hat{R}_{i+k}$, we will evaluate it from four dimensions, including: Fluency, Relevance, Completeness, and Proficiency. These dimensions are adopted as per suggestions from experts.

4 Experiments on CMB-Exam

4.1 Experimental Setup

Models We evaluate the following Chinese medical LLMs to compare their performance on CMB-Exam: HuatuoGPT-II (Chen et al., 2023a), ChatMed-Consult (Zhu and Wang, 2023), MedicalGPT (Xu, 2023), ChatGLM-Med (Wang et al., 2023b), DoctorGLM (Xiong et al., 2023), BianQue-2 (Chen et al., 2023b), Bentsao (Wang et al., 2023a), IvyGPT (Wang et al., 2023c), Sunsimiao (Xin Yan, 2023) and DISC-MedLLM (Bao et al., 2023). In addition to these specialized models, we also evaluate some publicly-available general-domain instruction-following model series, namely ChatGLM3-6B (Du et al., 2022), Baichuan2 (Baichuan, 2023), Qwen (Bai et al., 2023), Yi[2], Deepseek-llm (AI, 2023), Mistral (Jiang et al., 2023) and Internlm (Team, 2023). For closed source commercial models, we evaluate ChatGPT[3], GPT-4, ShunkunGPT, AntGLM-Med and JianpeiGPT. We also test the performance of DISC-MedLLM trained on CMB-train. Please refer to Appendix C for more details about models and training.

Decoding Hyperparameters For all the aforementioned open source models, we adopt their default hyper-parameters. Besides, to reduce the variance in generation, we adopt greedy decoding for all models on both CMB-Exam and CMB-Clin. And the `min_new_tokens` and `max_new_tokens` are set to 1 and 512, respectively, to avoid empty or lengthy answers on CMB-Exam.

Evaluation Details We evaluate the models in both answer-only and chain-of-thought (CoT) settings. We extract answers from model outputs using an empirically designed regular expression. Each extracted answer is compared to the solution and is deemed correct if and only if they are exactly matched. We adopt accuracy as our metric. All evaluation experiments and training experiments take around 1000 GPU-hours on 8 NVIDIA A800 80GB GPUs.

[2] https://github.com/01-ai/Yi
[3] gpt-3.5-turbo-16k-0613

现病史

病史摘要：周XX，男，25岁，1年前因车祸致小肠损伤，行部分小肠切除吻合术。术中切除小肠约40cm。术后病人恢复良好，未诉特殊不适。1个月前病人无明显诱因突然出现腹痛、腹胀、停止排气排便，至急诊就诊，诊断为肠梗阻，急诊行剖腹探查术。术中发现小肠与腹壁多处粘连带压迫小肠肠管，距屈氏韧带60cm以下至距回盲部约25cm间小肠缺血坏死。遂切除坏死肠管，行空肠、回盲部端侧吻合 。 术后至今病人每日大便10余次，为水样便，量较多，无脓血，无腹痛、腹胀。 **查体：**体温：36.9℃，血压：115/78mmHg，脉搏：96次/分。腹部平坦，全腹无压痛、反跳痛，无肌紧张。 **辅助检查：**大便常规提示：WBC 0-1/HPF，潜血（-）。**血常规提示：**WBC 5.5×109/L，Hb 102g/L，血小板计数：192×109/L。**血生化提示：**Na+：134mmol/L，K+：3.4mmol/L，Ca2+：2.42mmol/L，Mg2+：0.75mmol/L。近1个月体重下降约4kg，睡眠质量差。

Present Medical History

Summary of Medical History: Mr. XX, male, 25 years old, had a small intestine injury due to a car accident one year ago, underwent partial small intestine resection and anastomosis surgery. Approximately 40cm of the small intestine was removed during the operation. The patient recovered well postoperatively and did not report any specific discomfort. One month ago, he suddenly experienced abdominal pain, abdominal distension, and cessation of gas and bowel movements without apparent cause. He sought emergency medical attention and was diagnosed with intestinal obstruction. Emergency exploratory laparotomy revealed adhesions between the small intestine and the abdominal wall, compressing the small intestine, with ischemic necrosis approximately 60cm below the ligament of Treitz to about 25cm proximal to the ileocecal junction. The necrotic segment was removed, and an end-to-end anastomosis between the jejunum and the ileocecal region was performed. Since the surgery, the patient has been having more than 10 watery bowel movements per day, with a significant amount, no pus or blood, and no abdominal pain or distension. **Physical Examination:** Temperature: 36.9℃, Blood Pressure: 115/78mmHg, Pulse: 96 beats/min. The abdomen is flat, with no tenderness, rebound tenderness, or muscle tension throughout. **Ancillary Examinations:** Stool routine: WBC 0-1/HPF, occult blood (-). **Blood Routine:** WBC 5.5×109/L, Hb 102g/L, Platelet count: 192×109/L. **Blood Biochemistry:** Na+: 134mmol/L, K+: 3.4mmol/L, Ca2+: 2.42mmol/L, Mg2+: 0.75mmol/L. Weight loss of approximately 4kg in the past month, poor sleep quality.

问题1：该病人目前的诊断是什么？

Question 1: What is the current diagnosis of the patient?

参考答案：短肠综合征（急性期）。病人因车祸及术后粘连致肠坏死两次行小肠切除手术，小肠残留长度不足100cm，术后不足2个月，出现了大量腹泻、电解质紊乱及体重明显下降，符合短肠综合征急性期的临床表现。

Reference: Short Bowel Syndrome (acute phase). The patient underwent two small intestine resection surgeries due to a car accident and postoperative adhesions leading to intestinal necrosis. The remaining length of the small intestine is less than 100cm. Less than two months after surgery, the patient has developed significant diarrhea, electrolyte disturbances, and substantial weight loss, which are consistent with the clinical manifestations of the acute phase of Short Bowel Syndrome.

问题2：该病人的治疗方法是什么？

Question 2: What is the treatment method for this patient?

参考答案：①尽快给予全胃肠外营养支持，补充病人正常所需水份、钠钾镁等电解质、能量及营养物质；②给予抑制肠蠕动药物；③给予H2受体拮抗剂或质子泵抑制剂；④腹泻症状初步被控制后，给予少量等渗肠内营养。

Reference: ① Provide total parenteral nutrition support as soon as possible to supplement the patient's normal water, electrolytes such as sodium, potassium, and magnesium, energy, and nutrients; ② Administer drugs to inhibit intestinal motility; ③ Administer H2 receptor antagonists or proton pump inhibitors; ④ After the symptoms of diarrhea are initially controlled, provide a small amount of isotonic enteral nutrition.

Figure 1: Case of Short Bowel Syndrome from CMB-Clin. English translations are shown for better readability.

4.2 Benchmarking Results

We report the results in Table 3 and Table 4. There are several observations drawn from different aspects.

On general LLMs. As shown in Table 3, there is no general model that is particularly lacking in medical knowledge. Taking GPT-4 and ChatGPT as the watershed, Qwen-72B, Yi-34B, and Yi-6B have achieved accuracy exceeding GPT-4, and most models have achieved performance exceeding ChatGPT. Yi-6B achieves such good results with a smaller number of parameters is impressive, but it also reminds us of the possibility of data leakage. At the same time, we also noticed that Qwen-1.8B showed strong in-context learning capabilities. Considering its number of parameters, this is also an exciting discovery. Domestic general models have completed catching up with OpenAI in medical knowledge in chinese.

On medical LLMs. As shown in Table 4, the gap between medical models is obvious. HuatuoGPT-II surpasses all commercial models and ranks first, demonstrating its outstanding capabilities. At the same time, the commercial model is also significantly ahead of other open source medical LLMs except HuatuoGPT-II. Considering that it has been a long time since most medical LLMs were open sourced, we believe that the new generation of open source medical LLMs will quickly close the gap. After trained on CMB-train, DISC-MedLLM ranked second among open source medical models, indicating that the large amount of deterministic medical knowledge contained in multiple-choice questions in CMB-Train is very helpful for improving performance. How to efficiently inject the knowledge of multiple-choice questions into LLM is a promising task.

On different categories. LLMs show varied performance across clinical specialties. Specifically, scores for pharmacist-related questions tend to be lower, while those concerning nursing staff are typically higher. This difference might arise from the foundational knowledge nurses require, which is straightforward, compared to the intricate distinctions in drug names and indications pharmacists deal with. Despite these performance variations among specialties, the models exhibit a consistent trend, suggesting no inherent bias towards any particular domain.

On prompt strategies For the vast majority of domestic General LLM and Medical LLM, both the Few-shot and CoT strategies have little effect on improving model accuracy. The few-shot strategy has improved significantly for models such as Deepseek-llm, Mistral-7B, Mixtral-8x7B, which originally had limited support for Chinese, and smaller models such as Qwen-1.8B and Yi-6B. The CoT strategy even has negative effects on models such as Mistral and ChatGLM-Med, which have very low original accuracy. In CMB-Exam, for

Model	Open	Physician	Nurse	Pharmacist	Technician	Disciplines	Graduate Entrance Exam	Average
General Models								
Qwen-72B-Chat	✓	**78.55 (80.00)**	**83.56 (84.06)**	79.78 (**80.31**)	**77.92 (79.50)**	**68.26 (67.55)**	58.19 (**57.44**)	**74.38 (74.81)**
+ CoT		78.40 (76.15)	83.31 (81.69)	**80.13** (76.78)	77.08 (73.00)	67.69 (65.38)	**58.81** (55.50)	74.24 (71.42)
Yi-34B-Chat	✓	71.10 (72.95)	77.56 (80.44)	73.16 (74.03)	73.67 (76.92)	66.56 (67.31)	52.94 (55.63)	69.17 (71.22)
+ CoT		69.05 (58.45)	74.75 (63.13)	70.25 (60.06)	68.00 (57.08)	63.00 (56.31)	51.69 (45.44)	66.12 (56.75)
Yi-6B-Chat	✓	67.25 (68.70)	76.38 (77.06)	68.50 (69.38)	67.83 (68.00)	61.75 (62.44)	53.50 (55.88)	65.87 (66.91)
+ CoT		64.30 (59.95)	73.50 (67.38)	64.44 (61.03)	65.33 (56.67)	59.63 (55.25)	50.50 (48.75)	62.95 (58.17)
GPT-4	✗	59.90 (60.19)	69.31 (70.14)	52.19 (53.25)	61.50 (62.38)	59.69 (60.25)	54.19 (55.12)	59.46 (60.22)
+ CoT		58.15 (59.63)	70.31 (71.02)	53.09 (52.15)	62.34 (61.38)	60.69 (62.25)	52.30 (54.21)	59.45 (60.11)
Deepseek-llm-67B-Chat	✓	52.90 (59.15)	61.50 (65.19)	54.28 (59.22)	51.42 (55.25)	51.19 (55.63)	40.63 (38.88)	51.99 (55.55)
+ CoT		56.20 (51.80)	60.19 (60.25)	54.44 (53.69)	50.75 (47.58)	51.38 (51.63)	41.00 (38.50)	52.33 (50.58)
Baichuan2-13B-Chat	✓	49.55 (50.05)	56.75 (57.75)	49.41 (50.50)	50.08 (49.50)	48.25 (49.06)	39.18 (40.63)	48.87 (49.58)
+ CoT		48.90 (48.55)	57.25 (54.75)	49.88 (49.16)	49.33 (47.08)	46.88 (44.37)	38.06 (36.44)	48.38 (46.73)
Qwen-7B-Chat	✓	48.00 (49.45)	54.25 (55.13)	48.34 (47.94)	48.08 (49.25)	44.87 (45.00)	35.94 (36.56)	46.58 (47.22)
+ CoT		48.00 (45.65)	54.25 (52.63)	48.34 (47.28)	48.08 (43.08)	44.88 (44.06)	35.94 (36.19)	46.58 (44.82)
Baichuan2-7B-Chat	✓	42.55 (43.30)	51.75 (51.56)	44.59 (44.59)	45.50 (43.00)	43.00 (40.44)	32.56 (34.00)	43.33 (42.82)
+ CoT		43.55 (38.25)	51.06 (47.13)	44.72 (43.91)	43.17 (39.50)	42.69 (39.63)	32.44 (20.56)	42.94 (38.16)
ChatGLM3-6B	✓	42.55 (40.30)	47.31 (44.81)	39.56 (38.22)	41.08 (39.33)	37.44 (37.63)	32.06 (32.13)	40.00 (38.74)
+ CoT		38.90 (38.15)	45.38 (43.25)	38.19 (34.53)	38.08 (37.08)	36.12 (35.25)	30.13 (26.75)	37.80 (35.84)
ChatGPT	✗	40.75 (40.75)	45.69 (45.69)	36.59 (36.59)	40.08 (40.08)	37.94 (37.94)	28.81 (28.81)	38.31 (38.31)
+ CoT		17.75 (17.75)	19.94 (19.94)	16.00 (16.00)	20.25 (20.25)	19.25 (19.25)	16.19 (16.19)	18.23 (18 .23)
Internlm-Chat-20B	✓	39.35 (39.55)	45.44 (43.00)	38.53 (36.25)	37.92 (38.25)	38.12 (38.06)	29.63 (29.63)	38.17 (37.46)
+ CoT		39.60 (34.00)	44.44 (41.00)	36.41 (32.50)	40.08 (34.17)	37.88 (32.81)	30.88 (26.19)	38.21 (33.45)
Internlm-Chat-7B	✓	34.45 (32.55)	42.13 (36.81)	33.69 (32.41)	37.50 (35.00)	33.75 (31.06)	27.94 (26.94)	34.91 (32.46)
+ CoT		35.55 (34.70)	41.38 (38.31)	33.88 (32.41)	35.83 (35.42)	33.88 (32.19)	27.88 (25.50)	34.73 (33.09)
Mixtral-8x7B-32kseqlen	✓	34.50 (39.00)	42.00 (41.88)	25.06 (33.13)	30.83 (39.50)	31.81 (36.44)	22.25 (28.25)	31.07 (36.37)
+ CoT		34.50 (28.00)	42.00 (34.06)	25.06 (24.69)	30.83 (34.75)	31.81 (27.50)	22.25 (17.56)	31.07 (27.76)
Qwen-1.8B-Chat	✓	26.20 (44.15)	34.06 (50.63)	28.03 (39.78)	27.08 (39.25)	25.69 (36.56)	23.50 (33.75)	27.43 (40.69)
+ CoT		26.20 (30.95)	34.06 (41.50)	28.03 (32.25)	27.08 (28.00)	25.69 (27.81)	23.50 (28.00)	27.43 (31.42)
Mistral-7B-Instruct-v0.1	✓	23.75 (19.55)	22.19 (22.50)	20.97 (19.88)	25.83 (21.42)	21.94 (19.25)	18.88 (16.75)	22.26 (19.89)
+ CoT		21.90 (19.95)	23.06 (21.44)	20.97 (19.97)	23.08 (20.83)	21.81 (19.00)	15.56 (12.94)	19.02 (19.02)

Table 3: Accuracy in the answer-only and CoT settings across different categories for **general models**. Values in parentheses are the Three-shot accuracy.

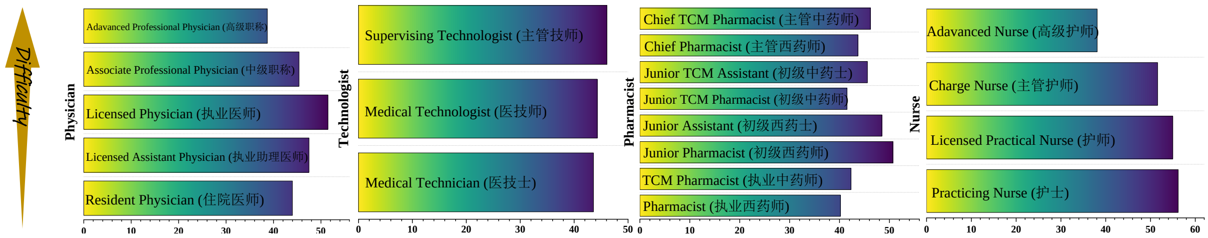

Figure 2: Accuracy across various clinical medicine fields at different career stages. The accuracies are the Zero-shot average values for all the models using direct response strategy. The difficulty increases from bottom to top.

problems that do not require reasoning, the CoT strategy may confuses correct information with irrelevant context, thereby reducing accuracy.

On the perceived difficulty As shown in Figure 2, the professional level continues to improve from bottom to top. Only the Nurse category meets expectations with accuracy decreases from bottom to top. For the Physician, Advanced Professional subcategory have the lowest accuracy and Resident Physician have the second lowest accuracy. After sample analysis, we found that the questions covered in the Resident Physician subcategory involve many uncommon details and knowledge, which increases the probability of hallucinations. For Technologist, it's interesting that the accuracy rate is completely opposite to expectations. We found that there are many questions focus on personnel management and communication in Supervising Technologist subcategory, which is indeed what LLMs are good at. For the Pharmacists, there is no obvious trending. But subcategories involving traditional Chinese medicine always have relative low accuracy, indicating that additional data on traditional Chinese medicine still needs to be supplemented.

5 Experiments on CMB-Clin

5.1 Experimental Setup

Prompt construction Every prompt comprises two components: a description that may (or may

Model	Open	Physician	Nurse	Pharmacist	Technician	Disciplines	Graduate Entrance Exam	Average
Commercial Medical Models								
JianPeiGPT	✗	73.60*	77.00*	72.84*	65.00*	70.13*	78.40*	72.84*
ShuKunGPT	✗	68.65*	71.44*	70.78*	61.92*	62.81*	51.06*	64.44*
AntGLM-Med	✗	62.85*	66.81*	60.06*	48.50*	54.69*	51.06*	55.00*
Open source Medical Models								
HuatuoGPT-II-34B (华佗II)	✓	75.65 (75.65)	82.31 (**82.31**)	**76.81 (77.12)**	76.17 (**74.12**)	**74.38 (74.38)**	**75.56** (75.56)	**76.82 (76.52)**
+ CoT		**76.13 (76.13)**	**83.15** (82.15)	**76.81** (76.81)	**77.12** (70.12)	71.22 (70.22)	**75.56 (76.12)**	76.67 (75.26)
HuatuoGPT-II-13B (华佗)	✓	67.85 (67.85)	66.12 (66.12)	64.91 (64.91)	62.00 (63.05)	61.94 (62.15)	53.69 (54.69)	62.75 (63.13)
+ CoT		68.02 (68.05)	65.32 (65.32)	65.12 (65.12)	63.01 (62.55)	62.01 (61.53)	54.60 (54.63)	63.01 (62.87)
HuatuoGPT-II-7B (华佗)	✓	64.55 (64.55)	63.75 (63.75)	61.06 (61.06)	56.25 (56.25)	56.63 (56.90)	51.81 (53.82)	59.00 (59.39)
+ CoT		65.12 (65.12)	64.33 (63.12)	60.05 (61.50)	57.12 (56.03)	56.63 (57.01)	51.81 (52.18)	59.18 (59.16)
DISC-MedLLM-13B	✓	42.25 (42.20)	46.88 (47.87)	38.44 (38.94)	38.83 (38.92)	40.75 (39.38)	31.44 (31.25)	39.76 (39.76)
+ CoT		41.85 (41.30)	47.19 (46.44)	38.97 (38.41)	39.17 (38.17)	40.31 (39.81)	31.37 (31.44)	39.78 (39.26)
IvyGPT-13B	✓	37.70 (37.34)	43.56 (43.56)	40.47 (41.25)	38.08 (39.06)	35.31 (36.31)	36.12 (37.15)	38.54 (39.11)
+ CoT		37.15 (38.23)	44.12 (45.12)	41.23 (42.33)	38.08 (39.12)	36.12 (37.20)	36.12 (36.88)	38.80 (39.81)
Sunsimiao-7B (孙思邈)	✓	38.75 (38.12)	44.37 (45.12)	38.81 (39.12)	38.33 (38.33)	37.50 (38.12)	33.31 (34.21)	38.51 (33.13)
+ CoT		39.12 (39.12)	45.12 (45.12)	38.81 (39.12)	38.33 (39.31)	37.50 (38.12)	34.12 (34.12)	38.84 (39.96)
MedicalGPT-7B	✓	26.40 (26.56)	30.94 (30.94)	24.72 (24.84)	27.17 (27.32)	25.44 (25.62)	21.50 (21.64)	26.03 (26.15)
+ CoT		24.80 (25.61)	27.19 (27.98)	23.09 (24.07)	24.58 (26.00)	23.75 (24.77)	21.06 (21.79)	24.08 (25.04)
ChatMed-Consult-7B	✓	20.20 (21.41)	22.31 (23.48)	20.59 (21.58)	22.67 (23.55)	20.38 (21.36)	17.44 (18.08)	20.60 (21.58)
+ CoT		19.40 (20.92)	21.69 (23.56)	20.00 (21.65)	22.83 (23.59)	18.88 (20.44)	18.56 (19.55)	20.23 (21.62)
ChatGLM-Med-7B	✓	21.75 (23.59)	22.06 (23.37)	21.84 (22.67)	21.00 (21.85)	18.44 (19.72)	17.50 (18.14)	20.43 (21.56)
+ CoT		15.55 (20.89)	16.25 (22.13)	17.34 (21.06)	16.33 (20.65)	12.63 (17.12)	12.56 (16.88)	15.11 (19.79)
Bentsao-7B (本草)	✓	21.55 (21.67)	19.94 (19.99)	20.94 (21.07)	22.75 (22.85)	19.56 (19.83)	16.81 (16.93)	20.26 (20.39)
+ CoT		21.00 (21.10)	20.56 (20.61)	20.66 (20.78)	22.17 (22.24)	19.25 (19.53)	16.44 (16.54)	20.01 (20.13)
BianQue-2 (扁鹊-2)	✓	4.90 (4.40)	4.19 (5.19)	4.28 (7.97)	3.58 (8.08)	3.31 (5.69)	3.25 (4.00)	3.92 (5.89)
+ CoT		7.85 (6.95)	6.63 (7.31)	7.34 (7.25)	8.33 (9.75)	6.63 (6.94)	5.94 (6.06)	7.12 (7.38)
DoctorGLM	✓	2.70 (0.10)	3.31 (0.38)	3.84 (0.34)	3.75 (0.50)	3.19 (0.37)	2.25 (0.81)	3.17 (0.42)
+ CoT		3.15 (2.35)	3.13 (2.50)	3.41 (3.28)	2.50 (1.17)	3.38 (3.06)	2.25 (3.88)	2.97 (2.71)
Models Trained by CMB-train								
DISC-MedLLM-13B (CMB-train)	✓	43.22 (43.22)	48.13 (47.56)	39.12 (40.23)	40.12 (45.12)	41.25 (42.25)	33.25 (33.75)	40.85 (42.02)
+ CoT		42.65 (43.65)	47.15 (48.13)	40.12 (41.22)	39.32 (40.12)	42.25 (41.58)	33.80 (34.80)	40.88 (41.58)

Table 4: Accuracy in the answer-only and CoT settings across different categories for **medical models**. Values in parentheses are the Three-shot accuracy. * means we only have the best score and the generation strategy is unknown.

Aspects	GPT-4	Yi-34B	Qwen-72B	ChatGPT	Baichuan2-13B	ChatGLM3-6B	Internlm-20B	Deepseekllm-67B	Mixtral-8x7B
Fluency	4.95	4.99	4.96	4.97	4.93	4.92	4.9	4.78	2.53
Relevance	4.71	4.69	4.58	4.49	4.41	4.11	3.91	4.04	2.28
Completeness	4.35	4.34	4.12	4.12	4.03	3.74	3.25	2.62	1.54
Proficiency	4.66	4.64	4.55	4.53	4.36	4.23	4.14	4.16	3.04
Average	4.67	4.67	4.55	4.53	4.43	4.25	4.05	3.9	2.35

Table 5: Results of CMB-Clin on Automatic Evaluation using GPT-4 for General Models.

not) encompass conversation history D_i, and the question Q_i. To integrate the conversation history into the description, we prepend the appropriate roles to each question and reference.

Expert and Automatic Evaluation To prove the validity of our evaluation, we engage three annotators with professional medical knowledge to evaluate on a randomly selected subset of 320 responses of 11 models from different tiers. Equipped with a reference solution, they score each response across four aspects — Fluency, Relevance, Completeness, and Medical Proficiency — using a grading scale from 1 to 5. The user interface is shown in Appendix D.1.1. To enhance efficiency and reduce expert evaluation costs, we leverage GPT-4 to assess the responses of all models, adhering to the same guidelines as those used in expert evaluations. The prompt template for the automatic evaluation is detailed in Appendix D.1.2.

5.2 Benchmarking Results

On general LLMs As shown in Table 5, except for Deepseekllm-67B and Mixtral-8x7B, which have insufficient support for Chinese models, the other General LLMs have shown strong dialogue capabilities based on complex medical records. Taking GPT-4 and ChatGPT as the dividing line, Yi-34B has achieved the same medical dialogue capability as GPT-4. Qwen-72B is weaker than GPT-4 but stronger than ChatGPT, and the remaining models are all weaker than ChatGPT. Compared with

Aspects	HuatuoGPT-II-34B	BianQue-2	DISC-MedLLM	ChatMed-Consult	MedicalGPT	DISC-MedLLM-Train	DoctorGLM	Bentsao	ChatGLM-Med
Fluency	4.96	4.86	4.82	4.88	4.48	4.57	4.74	3.88	3.55
Relevance	4.61	3.52	3.24	3.08	2.64	2.52	2.00	2.05	1.97
Completeness	4.31	3.02	2.75	2.67	2.19	1.89	1.65	1.71	1.61
Proficiency	4.53	3.60	3.51	3.30	2.89	3.19	2.30	2.58	2.37
Average	4.60	3.75	3.58	3.48	3.05	3.04	2.67	2.55	2.38

Table 6: Results of CMB-Clin on Automatic Evaluation using GPT-4 for Medical Models.

their strong performance in CMB-Exam, domestic General LLMs still lag behind OpenAI in CMB-Clin, which is closer to real scenarios. Except for the Yi LLMs, the ability of other domestic LLMs to solve real problems does not match their ability to answer multiple-choice questions, suggesting that they may have been specially strengthened for multiple-choice questions. Such model iteration direction actually deviates from actual needs. During the iteration process, we recommend using both CMB-Exam and CMB-Clin for model capability awareness.

On medical LLMs As shown in Table 6, the overall dialogue ability of Medical LLMs is lower than that of General LLMs. Although the three models of DoctorGLM, Bentsao, and ChatGLM-Med all claim to be optimized for consultation, the actual results show that their conversational capabilities have not been enhanced. It is worth noting that although BianQue-2 performed poorly in CMB-Exam, it performed well in CMB-Clin, indicating that it just lacks the ability to do multiple-choice questions and follow instructions. Although HuatuoGPT-II-34B surpasses GPT-4 in CMB-Exam, it still lags behind GPT-4 and is even lower than its base model Yi-34B in CMB-Clin, indicating that multiple rounds of dialogue data need to be added during its training process.It is noted that performance of DISC-MedLLM trained on CMB-train drops significantly on CMB-Clin, indicating the need to add other data or reconstruct multiple-choice questions in the form of dialogues.

On different Metrics Regarding the Fluency indicator, there is not much difference between General LLMs with most LLM above 4.9, but there are still many Medical LLMs models below 4.5, indicating a lack of basic dialogue capabilities. Relevance, Completeness and Proficiency are all highly differentiated indicators, among which Completeness has the lowest average value, indicating that for medical record consultation scenarios, the completeness of the dialogue and obtaining complete information are the most difficult task.

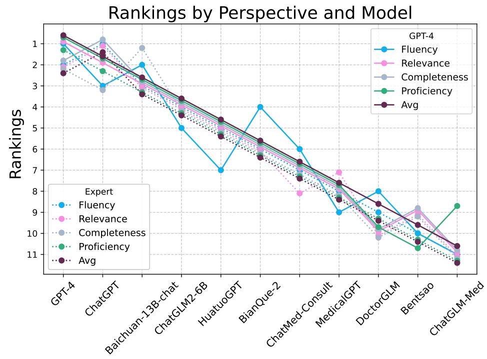

Figure 3: Rankings by perspective and model. Dashed lines and solid lines are the resulted rankings from expert and ChatGPT evaluation, respectively. For visual clarity, each line is shifted vertically for a small value. A model is better if it has a smaller ranking (a higher position) on the vertical axis.

5.3 Agreements between Automatic and Expert Evaluation

Figure 3 demonstrates a strong agreement of resulted rankings between GPT-4 and expert evaluation, with the spearman correlation of rankings being 0.93. The rankings agree with each other except for a flip for GPT-4 and ChatGPT (dashed and solid brown lines are parallel, except for a flip at GPT-4 and ChatGPT). We also provide a more fine-grained alignment analysis in Appendix D.1.4. The results indicate that the automatic evaluation is highly aligned with expert evaluation.

6 Conclusion

Although LLMs have potential in the realm of medicine, their accurate evaluation remains pivotal for real-world applications. The introduction of the CMB benchmark, tailored to the local cultural environment in China, gives a more contextualized and comprehensive evaluation benchmark. Although not framed as a competitive leaderboard, it serves as a crucial tool for tracking LLM progress in medical domains. This might pave the way for a broader and more effective utilization of LLMs in China's medical landscape.

Ethical Statement

The data utilized in this study primarily originate from publicly accessible mock examination questions, coursework exercises, and summations of commonly misunderstood examination questions. A portion of these items are sourced from the Chinese Medical Question Database[4], from whom we received explicit permission and support to include their questions in our evaluation.

Limitations

The limitations of our study are twofold. Firstly, while our benchmark encompasses various subjects in the Chinese medical domain, there remain numerous subjects that necessitate multi-modal capabilities for addressing real-world issues. Secondly, within the CMB-Clin section, we standardized the model evaluation method. However, in the real world, diverse medical conditions may require distinct evaluation criteria. Therefore, we advocate the adoption of disease-specific criteria for assessing model performance.

[4] https://www.medtiku.com/

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at TREC 2017 liveqa. In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, volume 500-324 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R. Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All - Proceedings of the 17th World Congress on Medical and Health Informatics, Lyon, France, 25-30 August 2019*, volume 264 of *Studies in Health Technology and Informatics*, pages 25–29. IOS Press.

DeepSeek AI. 2023. Deepseek llm: Let there be answers. https://github.com/deepseek-ai/DeepSeek-LLM.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation.

Junying Chen, Xidong Wang, Anningzhe Gao, Feng Jiang, Shunian Chen, Hongbo Zhang, Dingjie Song, Wenya Xie, Chuyi Kong, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023a. Huatuogpt-ii, one-stage training for medical adaption of llms.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Zhipei Xu, Kai Fang, Sihang Li, Junhong Wang, and Xiangmin Xu. 2023b. Bianque-1.0: Improving the "question" ability of medical chat model through finetuning with hybrid instructions and multi-turn doctor qa datasets. *github*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. 2021a. Measuring coding challenge competence with APPS. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou andchen2023bianque1 Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021c. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. 2023. Benchmarking large language models on cmexam - A comprehensive chinese medical exam dataset. *CoRR*, abs/2306.03030.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer J. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2357–2368. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 2019 Workshop on Biomedical Natural Language Processing (BioNLP 2019)*.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J. Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. ARB: advanced reasoning benchmark for large language models. *CoRR*, abs/2307.13692.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, Gregory S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle K. Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Simon Suster and Walter Daelemans. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1551–1563. Association for Computational Linguistics.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge.

Haochun Wang, Chi Liu, Sendong Zhao Zhao, Bing Qin, and Ting Liu. 2023b. Chatglm-med: 基于中文医学知识的chatglm模型微调. https://github.com/SCIR-HI/Med-ChatGLM.

Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, Chongpei Chen, Ruyi Gan, and Jiaxing Zhang. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Rongsheng Wang, Yaofei Duan, ChanTong Lam, Jiexi Chen, Jiangsheng Xu, Haoming Chen, Xiaohong Liu, Patrick Cheong-Iao Pang, and Tao Tan. 2023c. Ivygpt: Interactive chinese pathway language model in medical domain. *arXiv preprint arXiv:2307.10512*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguistics*, 6:287–302.

Dong Xue* Xin Yan. 2023. Sunsimiao: Chinese medicine llm. https://github.com/thomas-yanxin/Sunsimiao.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023a. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3840–3849. Association for Computational Linguistics.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2472–2482. ACM.

Wei Zhu and Xiaoling Wang. 2023. Chatmed: A chinese medical large language model. https://github.com/michael-wzhu/ChatMed.

A Related work

A.1 Medical Benchmark

Medical benchmarks have evolved to broadly encompass two types of tasks based on the capabilities of the models they seek to probe: objective tasks and subjective tasks. The former typically assumes the form of multiple-choice questions (Welbl et al., 2018; Pal et al., 2022; Hendrycks et al., 2021b; Singhal et al., 2022), information retrieval (Abacha et al., 2017; Zhu et al., 2019; Abacha et al., 2019), and cloze-style reading comprehension (Suster and Daelemans, 2018; Pampari et al., 2018; Zhu et al., 2020), which serve to evaluate a model's medical knowledge with unbiased accuracy. Sources for these tasks range from medical textbooks and exams to case reports (Suster and Daelemans, 2018), Wikipedia (Welbl et al., 2018), and medical practices (Hendrycks et al., 2021b; Pal et al., 2022). In contrast, subjective tasks involve open-ended text generation constructed directly from consumer queries and doctor responses, often sourced from online medical forums. The task typically demands models to generate consumer-oriented replies (Singhal et al., 2022; Li et al., 2023) or explanations for multiple-choice questions (Liu et al., 2023). As of now, there are relatively few open-ended text generation question-answering tasks that specifically center around providing consultation based on diagnostic reports.

Few existing benchmark datasets encapsulate both task types, with MultiMedQA (Singhal et al., 2022) and CMExam (Liu et al., 2023) sharing the closest resemblance to our work. Differing from prior work, our dataset exceeds in size and includes questions not only from the Chinese National Medical Licensing Examination but also from various authoritative medical textbooks.

A.2 Other Benchmarks of Large Language Models

The explosive growth in the number and capability of LLMs has led to a multitude of works aiming to discern their true capacity, evaluating both their general and specific abilities. General ability benchmarks include comprehensive test suites, each targeting different aspects of LLM's proficiency, ranging from handling multi-turn dialogues (Zheng et al., 2023) to gauging language comprehension and reasoning abilities (Srivastava et al., 2022; Zhang et al., 2023a; Zhong et al., 2023).

In terms of specific abilities, several benchmarks, apart from those related to medicine, aim to evaluate different capabilities of models. ARB (Sawada et al., 2023) was introduced to assess LLMs' performance in high-level reasoning tasks across multiple domains. C-Eval (Huang et al., 2023) serves as the first comprehensive benchmark to evaluate the advanced knowledge and reasoning abilities of Chinese-based models. Gaokao (Zhang et al., 2023b), MATH (Hendrycks et al., 2021c), and APPS (Hendrycks et al., 2021a) focus on assessing LLM proficiency in complex, context-specific tasks, and code generation, respectively.

B Dataset

Table 9, 10, 11 present a detailed directory structure of CMB-Exam. Initially, the organization is based on clinical professions and the exams commonly undertaken by these professionals, divided into six primary sections. Upon this foundation, each section is further categorized based on career progression and examination subjects. Within each sub-category, we have meticulously classified according to specific departments or courses.

C Details of Evaluated Models

In this section, we introduce and detail the models utilized in our evaluation. These models fall under three primary categories: 12 Chinese medical LLMs, 5 proprietary LLMs, and 13 publicly-available general-domain LLMs.

Chinese medical LLMs:

- HuatuoGPT-II: HuatuoGPT2 employs an innovative domain adaptation method to significantly boost its medical knowledge and dialogue proficiency. It showcases state-of-the-art performance in several medical benchmarks, especially surpassing GPT-4 in expert evaluations and the fresh medical licensing exams. The number of parameters of the model is 7B, 13B, and 34B.

- BianQue: It enhances its questioning ability by asking patients for more information to solve the issue that patients may not reveal all information in a single-turn conversation.

- ChatMed-Consult: It is built upon Chinese LLaMA (Cui et al., 2023) using real-world questions and synthetic responses from ChatGPT.

慢性呼吸衰竭患者机械通气的指征，错误的是 ___
Indications for mechanical ventilation in patients with chronic respiratory failure, the incorrect one is ____

A. PaCO2进行性升高
Progressive increase in PaCO2

B. PaCO2升高达到80mmHg以上
PaCO2 increases to above 80mmHg

C. 严重的低氧血症，积极氧疗后，PaO2仍小于40mmHg
Severe hypoxemia, after aggressive oxygen therapy, PaO2 is still less than 40mmHg

D. 呼吸频率超过35次/分
Respiratory rate exceeds 35 times/minute

E. 并发肺性脑病
Complicated with pulmonary encephalopathy

答案：B
Answer: B

解释：慢性呼吸衰竭是在原有肺部疾病基础上发生的，最常见病因为COPD，早期可表现为Ⅰ型呼吸衰竭，随着病情逐渐加重，肺功能愈来愈差，可表现为Ⅱ型呼吸衰竭。慢性呼吸衰竭稳定期，虽PaCO2降低和PaCO2升高，但患者通过代偿和治疗，可稳定在一定范围内，患者仍能从事一般的工作或日常生活活动。一旦由于呼吸道感染加重或其他诱因，可表现为PaCO2明显下降，PaCO2显著升高，此时可称为慢性呼吸衰竭的急性发作，这是我国临床上最常见的慢性呼吸衰竭类型。长期家庭氧疗应在极重度慢阻肺患者中应用，具体指征：①PaO2小于等于55mmHg或SaO2小于等于88%，有或无高碳酸血症；②PaO2为55～60mmHg或SaO2为0.55。长期家庭氧疗一般是经鼻导管吸入氧气，流量1.0～2.0L/分，每日吸氧持续时间大于15小时。故选B。

Explanation: Chronic respiratory failure occurs on the basis of existing lung diseases, with the most common cause being COPD. Early stages may manifest as Type I respiratory failure. As the condition progressively worsens and lung function deteriorates, it can manifest as Type II respiratory failure. During the stable period of chronic respiratory failure, although PaCO2 decreases and PaCO2 increases, patients can stabilize within a certain range through compensation and treatment, and they can still engage in general work or daily life activities. Once exacerbated due to respiratory infections or other triggers, it can manifest as a significant drop in PaCO2 and a significant increase in PaCO2, which can be referred to as an acute exacerbation of chronic respiratory failure, the most common type of chronic respiratory failure in our country's clinical practice. Long-term home oxygen therapy should be applied in patients with severe COPD, with specific indications: ①PaO2 less than or equal to 55mmHg or SaO2 less than or equal to 88%, with or without hypercapnia; ②PaO2 is 55～60mmHg or SaO2 is 0.55. Long-term home oxygen therapy is generally inhaled through a nasal cannula, with a flow rate of 1.0～2.0L/min, and the duration of oxygen inhalation is more than 15 hours per day. Therefore, the answer is B.

Figure 4: An development example with explanations in CMB-Exam. English translations are shown for better readability.

- MedicalGPT: It is based on Ziya-LLaMa (Wang et al., 2022) and adopts a four-stage training recipe, including continued pre-training, supervised fine-tuning, reward modeling, reinforcement learning.

- ChatGLM-Med: It is finetuned on ChatGLM-6B (Du et al., 2022) using instruction tuning data, which are built upon CMeKG[5].

- Bentsao: It is finetuned on LLaMa-7B (Touvron et al., 2023) using the same data as ChatGLM-Med.

- DISC-Med: DISC-MedLLM is a large model in the medical field specially designed for medical and health conversational scenarios.

- DISC-MedLLM-13B (CMB-train): The model after fine-tuning DISC-MedLLM on CMB-Train. ALL of above large language models are fine-tuned for 2 epoch on the full training set with a batch size of 32, with a learning rate of 105 using Adam. The warm-up rate of cosine scheduling is set to 0.03.

- IvyGPT: An LLM based on LLaMA that is trained and fine-tuned with high-quality medical question-answer (QA) instances and Reinforcement Learning from Human Feedback (RLHF).

- Sunsimiao: Sunsimiao is fine-tuned from Baichuan-7B and ChatGLM-6B series on 100,000-level high-quality Chinese medical data.

- DoctorGLM: Based on the Chinese consultation model of ChatGLM-6B, it is fine-tuned on a single A100 80G in 13 hours.

Proprietary models:

- ChatGPT: Developed by OpenAI, ChatGPT, rooted in the GPT-3.5 architecture, excels in both understanding and generating natural language.

- GPT-4: Another offering from OpenAI, GPT-4 employs deep learning techniques to elevate natural language processing capabilities, showcasing remarkable advancements across diverse tasks.

- JianPeiGPT:A Medical LLM developed by the Pecking Doctor team of Hangzhou Jianpei Technology Co., Ltd [6]. The evaluation results were submitted and made public on December 15, 2023 after the opening of CMB.

[5]https://github.com/king-yyf/CMeKG_tools

[6]http://www.jianpeicn.com/

- ShukunGPT: A Medical LLM developed by Shukun Technology[7]. The evaluation results were submitted and made public on October 23, 2023 after the opening of CMB.

- AntGLM-Med: A Medical LLM developed by the algorithm research group from AntGroup[8]. The evaluation results were submitted and made public on December 23, 2023 after the opening of CMB.

Publicly-available general-domain LLMs:

- ChatGLM-3: The third version of ChatGLM, which is an open source, bilingual dialogue language model.

- Baichuan2-chat: An advanced variant of Baichuan-13B model, focuses on dialogue tasks, boasting 13 billion parameters for efficient and effective conversation generation. The number of parameters of the model is 7B and 13B.

- Qwen-Chat: Tongyi Qianwen large model series developed by Alibaba Cloud based on Transformer, which is trained on extremely large-scale pre-training data. The number of parameters of the model is 1.8B, 7B and 72B.

- Yi: Large language models trained from scratch by developers at 01.AI[9]. The number of parameters of the model is 6B and 34B.

- Deepseek-llm-67B-Chat: An advanced language model comprising 67 billion parameters. It has been trained from scratch on a vast dataset of 2 trillion tokens in both English and Chinese.

- Internlm-Chat: It's pre-trained on over 2.3T Tokens containing high-quality English, Chinese, and code data. Chat version has undergone SFT and RLHF training, enabling it to better and more securely meet users' needs. The number of parameters of the model is 7B and 20B.

- Mistral: A 7B dense Transformer, fast-deployed and easily customisable. Small, yet powerful for a variety of use cases. Supports English and code, and a 8k context window.

- Mixtral-8x7B-32kseqlen: A 7B sparse Mixture-of-Experts model with stronger capabilities than Mistral 7B. Uses 12B active parameters out of 45B total. Supports multiple languages, code and 32k context window.

D Experiment Details

D.1 CMB-Clin

D.1.1 Screenshot of Human Evaluation UI

We show the screenshot of human evaluation UI in Figure 6 and Figure 7. We split the screenshot into two figures for better visual clarity.

D.1.2 Prompts for Automatic Evaluation

The prompt for automatic evaluation contains task instructions, metrics, criteria, and placeholders for information to be evaluated. It is designed based on the suggestion of experts and used by both ChatGPT and GPT-4.

```
You are an AI evaluator specializing in
    assessing the quality of answers
    provided by other language models.
    Your primary goal is to rate the
    answers based on their fluency,
    relevance, completeness, proficiency
     in medicine. Use the following
    scales to evaluate each criterion:
Fluency:
1: Completely broken and unreadable
    sentence pieces
2: Mostly broken with few readable
    tokens
3: Moderately fluent but with limited
    vocabulary
4: Mostly coherent in expressing complex
     subjects
5: Human-level fluency

Relevance:
1: Completely unrelated to the question
2: Some relation to the question, but
    mostly off-topic
3: Relevant, but lacking focus or key
    details
4: Highly relevant, addressing the main
    aspects of the question
5: Directly relevant and precisely
    targeted to the question

Completeness:
1: Extremely incomplete
2: Almost incomplete with limited
    information
3: Moderate completeness with some
    information
4: Mostly complete with most of the
    information displayed
5: Fully complete with all information
    presented

Proficiency in medicine:
```

[7]

[8] https://www.antgroup.com/en

[9] https://www.lingyiwanwu.com/

```
1: Using plain languages with no medical
     terminology.
2: Equipped with some medical knowledge
    but lacking in-depth details
3: Conveying moderately complex medical
    information with clarity
4: Showing solid grasp of medical
    terminology but having some minor
    mistakes in detail
5: Fully correct in all presented
    medical knowledge

You will be provided with the following
    information:
- a description
- a conversation based on the
    description (optional)
- a question based on the description
    and conversation
- the solution to the question
- a model's answer to the question

[description]
{description}
[end of description]

[conversation]
{history}
[end of conversation]

[question]
{question}
[end of question]

[solution]
{solution}
[end of solution]

[answer]
{answer}
[end of answer]
Make sure to provide your evaluation
    results in JSON format and ONLY the
    JSON, with separate ratings for each
     of the mentioned criteria as in the
     following example:
{`fluency': 3, `relevance': 3, `
    completeness': 3, `proficiency': 3}
```

Settings	Original	T-0.2	T-0.6	T-1.0	T-1.5
Original	1.00	0.95	0.90	0.87	0.87
T-0.2	0.95	1.00	0.98	0.88	0.88
T-0.6	0.90	0.98	1.00	0.90	0.90
T-1.0	0.87	0.88	0.90	1.00	1.00
T-1.5	0.87	0.88	0.90	1.00	1.00

Table 7: Pairwise Spearman correlations between results under different decoding temperatures. Original: results of greedy decoding (temperature 0). T-x: results of using nucleus sampling under temperature x.

D.1.3 Results of Expert evaluation on CMB-Clin

320 model responses are randomly sampled for this experiment due to a large number of answers to be evaluated and limited expert resources. We present the detailed results of expert evaluation in Table 8.

D.1.4 Agreement of Expert and GPT-4 Evaluation

Figure 5 shows the agreement between expert and GPT-4 evaluation on each perspective. The pearson correlations are all above 0.71, indicating a strong linear correlation between the two evaluation approaches.

D.1.5 Pairwise Correlation of Rankings under Different Temperatures

We evaluate the results generated under each setting (*i.e.,* under different temperatures) using ChatGPT. Then for each setting, we obtain a ranking for all models. We then calculate the pairwise spearman correlation between all sets of rankings. The results are summarized in Table 7.

Models	Fluency	Relevance	Completeness	Proficiency	Avg.
ChatGPT	**4.93**	**4.65**	**4.22**	4.34	**4.53**
GPT-4	4.88	4.61	4.20	**4.39**	4.52
Baichuan-13B-chat	4.79	4.29	**4.22**	4.30	4.40
ChatGLM2-6B	4.77	4.06	3.96	3.99	4.20
HuatuoGPT	4.70	3.89	3.69	3.81	4.02
BianQue-2	4.44	3.50	3.30	3.43	3.67
ChatMed-Consult	4.26	3.39	3.16	3.27	3.52
MedicalGPT	4.21	3.40	3.09	3.10	3.45
DoctorGLM	3.74	2.46	2.35	2.30	2.71
Bentsao	3.52	2.62	2.36	2.30	2.70
ChatGLM-Med	2.92	2.23	1.98	1.92	2.26

Table 8: Results of *expert* evaluation on CMB-Clin. *Avg.* are the averaged scores of each model over all perspectives. Models are arranged in descending order of *Avg.*

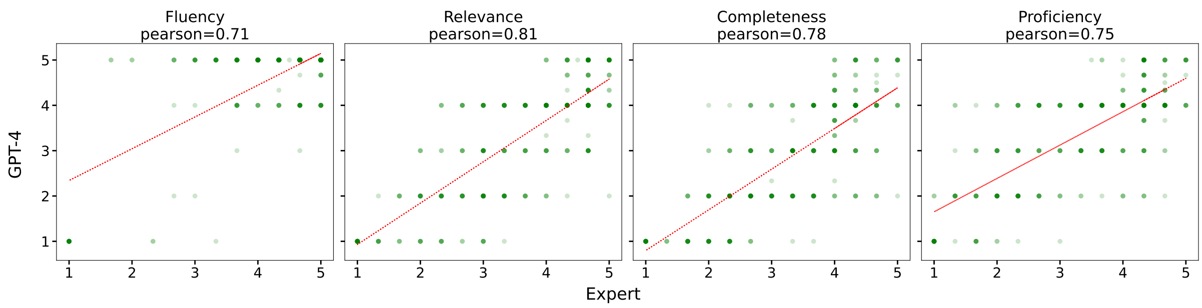

Figure 5: Correlation of expert and automatic evaluation on CMB-Clin of each perspective with pearson correlation. The four plots show correlations in fluency, relevance, completeness and proficiency in medicine, respectively. Each plot consists of 320 data points with many overlapped. The darker a point is, the more overlapped data there are at that position. Each expert score is averaged over the three expert annotators.

Category	Subcategory	Subject	# Questions
Physician	Resident Physician	Clinical Pathology	1124
		Oral	1074
		Otolaryngology	952
		Rehabilitation Medicine	461
		Ophthalmology	951
		Neurology	791
		Orthopedics	939
		Anesthesiology	907
		Pediatrics	749
		Dermatology	977
		Psychiatry	903
		General Practice	712
		Medical Imaging	964
		Internal Medicine	752
		Ultrasound	430
		Surgery	829
		Obstetrics and Gynecology	800
		Pediatric Surgery	296
	Licensed Assistant Physician	Integrated Chinese and Western Medicine	3441
		Clinical	5364
		Chinese Medicine	3454
		Public Health	2067
		Oral	1090
	Licensed Physician	Chinese Medicine	4490
		Public Health	4085
		Clinical	10241
		Oral	1505
		Integrated Chinese and Western Medicine	5320
	Associate Professional Physician	General Medicine	3492
		Internal Oral	858
		Orthopedics	894
		Chinese Internal Medicine	2896
		Surgery	5071
		Ultrasound Medicine	2218
		Dermatology and Venereology	1158
		Otolaryngology	983
		Internal Medicine	5671
		Infectious Diseases	600
		Obstetrics and Gynecology	2641
		Cardiovascular Internal Medicine and Respiratory Internal Medicine	617
		Oncology	942
		Acupuncture Attending in TCM	1169
		Pathology	1642
		Preventive Medicine	2817
		Pediatrics	3773
		Psychotherapy	1393
		Radiology	2401
		Psychiatry	754
		Oral Restoration	1183
		Dermatology	909
		Digestive Internal Medicine	160
		Rehabilitation Medicine	630
		Infectious Disease	861
		Nuclear Medicine	1250
		Oral Medicine	862
		Integrated Chinese and Western Internal Medicine	1101
		Ophthalmology	988
		Anesthesiology	923
		Hospital Infection	827
		Nutrition	1009
		Tuberculosis	58
		Critical Care Medicine	579
		Psychological Counselor	495
		Pain Medicine	884
		Neurology	126
		Orthodontics	578
		Oral and Maxillofacial Surgery	367
		Plastic Surgery	187
		Nephrology	81
		Rheumatology and Clinical Immunology	37
		Occupational Disease	54
	Advanced Professional Physicians	Respiratory InternalMedicine	1522
		Orthopedics	1245
		Endocrinology	1326
		Cardiology	1604
		Digestive Internal Medicine	1577
		General Surgery Senior	1850
		Gynecology and Obstetrics	3249
		General Internal Medicine	607
		General Practice	74
		Pediatrics	65

Table 9: Catalog Structure of Physician

Category	Subcategory	Subject	# Questions
Undergraduate Disciplines	Foudamental Medicine	Pathophysiology	1455
		Medical Psychology	932
		Biochemistry and MolecularBiology	2402
		Cell Biology	1399
		Medical Immunology	2485
		Pathology	2786
		Medical Genetics	1369
		Parasitology	806
		Systematic Anatomy	1967
		Bioinformatics	185
		Physiology	2306
		Pharmacology	2424
		Medical Microbiology	1342
		Local Anatomy	489
		Histology and Embryology	1398
		Human Parasitology	766
		Medical Statistics	198
	Clinical Medicine	Medical Imaging	1858
		Radiology	541
		Experimental Diagnostic Medicine	548
		Neurology	1163
		Surgery	2164
		Dermatology and Venereology	2168
		Pediatrics	3760
		Nuclear Medicine	1383
		Physical Diagnosis	621
		Dental Pulp Disease	346
		Basic Nursing	978
		Diagnostics	103
		Ultrasonic Medicine	192
		Oral Care	263
		Evidence-Based Medicine	95
		Fundamental Nursing	393
		Epidemiology	864
		Oral Tissue Pathology	387
		Infectious Disease	287
		Oral Anatomy and Physiology	362
		Anesthesiology	606
		Interventional Radiology	81
	TCM and Chinese Herbal Medicine	Preventive Medicine	1926
		Hygiene	1316
		Medical Ethics	500
	Preventive Medicine and Public Health	TCM Ophthalmology	915
		Essential Prescriptions Worth a Thousand Gold	1051
		Basic Theories of TCM	2706
		TCM Diagnosis	2036
		TCM	1921
		Warm Disease Theory	1088
		History of Chinese Medicine	662
		TCM Internal Medicine	1738
		TCM Pediatrics	694
		Treatise on Cold Pathogenic Diseases	1390
		Lecture on Inner Canon	456

Table 10: Catalog Structure of Undergraduate Disciplines

Category	Subcategory	Subject	# Questions
Nurse	Practicing Nurse	Practicing Nurse	3303
	Licensed Practical Nurse	Licensed Practical Nurse	4223
	Charge Nurse	Pediatric	905
		Internal Medicine	958
		Charge Nurse	4558
		Surgery	341
		Obstetrics and Gynecology	755
	Advanced Practice Nurse	Advanced Practice Nurse	1876
Technician	Medical Technician	Rehabilitation Medicine Therapy	1752
		Radiology	1033
		Inspection	1166
		Oncology	1086
	Medical Technologist	Rehabilitation Medicine Therapy	1739
		Oncology	1538
		Radiology	1337
		Inspection	1458
	Supervising Technologist	Radiation Therapy for Oncology	1701
		Ultrasonic Medicine	145
		Blood Transfusion Technology	2199
		Microbiological Inspection	704
		Radiology	1428
		Pathology	2407
		Physical and Chemical Inspection	783
		Clinical Medicine Inspection	1378
		Medical Record Information	1331
		Nuclear Medicine	1275
		Electrocardiology	1021
		Disinfection Technology	575
		Rehabilitation Medicine and Treatment	948
Graduate Entrance Exam	Nursing	Surgical Nursing	1112
		Basic Nursing	902
	Political Science	Political Science	1514
	Integrated Western Medicine	Integrated Western Medicine	8913
	Integrated TCM	Integrated TCM	3924
Pharmacist	Licensed Pharmacist	Licensed Pharmacist	8248
	Licensed TCM Pharmacist	Licensed TCM Pharmacist	4460
	Junior Pharmacist	Junior Pharmacist	2720
	Junior Pharmacist Assistant	Junior Pharmacist Assistant	3705
	Junior TCM Pharmacist	Junior TCM Pharmacist Assistant	3502
	Junior TCM Pharmacist	Junior TCM Pharmacist Assistant	4017
	Chief Pharmacist	Chief Pharmacist	3403
	Chief TCM Pharmacist	Chief TCM Pharmacist	3299

Table 11: Catalog Structure of Nurse, Technician, Graduate Entrance Exam and Pharmacist

用户名，按回车键提交

欢迎 user

测评细则（点击此处折叠/展开，请在开始评分前阅读） ▼

打分标准

流畅性

1分：完全破碎且无法阅读的句子片段
2分：大部分破碎，只有少量可读的词
3分：有一定流利度，但词汇有限
4分：在表达复杂主题方面基本上是流畅的
5分：人类水平流利度

相关性

1分：与问题完全无关
2分：与问题有一定关系，但主要是离题的
3分：相关，但缺乏重点或关键细节
4分：高度相关，解决了大部分问题
5分：直接相关，准确地针对了问题

完整性：

1分：极度不完整
2分：几乎不完整，信息有限
3分：有一定的完整性，有一些信息
4分：大部分信息都已显示
5分：所有信息都已呈现

医学知识专业性：

1分：使用简单明了的语言，没有医学术语
2分：具有一些医学知识，但缺乏深入细节
3分：清晰地传达了一定的复杂医学信息
4分：对医学术语有扎实的认识，但有些细节错误
5分：在所有呈现的医学知识上都是完全正确的

界面组成：

病史描述（左上）

包含病人基本信息、检查结果等。

对话（左下，可能为空）

包含与病史描述有关的问答对，作为额外信息提供给模型。

题目 （右上）

模型需要回答的问题。

模型回答 （右中）

模型根据**病史描述、历史对话、问题**给出的回答。

参考答案 （右下）

参考答案，作为对模型回答打分的依据。

测评

请阅读**病史描述、对话、题目**，对比**参考答案**，对**模型回答**进行打分。只有完成当前题目的4个维度的评分后，才能进入下一题。

本测评细则可折叠，评分过程中可随时再次查看/折叠。

Figure 6: The guideline for human evaluation and the introduction to components of user interface (in Chinese). Note that Figure 6 precedes Figure 7 in the same webpage.

Figure 7: The user interface for scoring an answer (in Chinese). Note that Figure 7 follows Figure 6 in the same webpage.