EMPEC: A Comprehensive Benchmark for Evaluating Large Language Models Across Diverse Healthcare Professions

Zheheng Luo[◇*] **Chenhan Yuan**[◇*] **Qianqian Xie**[†♠]
Sophia Ananiadou[◇]
[◇]The University of Manchester [†]School of Artificial Intelligence, Wuhan University

Abstract

Recent advancements in Large Language Models (LLMs) show their potential in accurately answering biomedical questions, yet current healthcare benchmarks primarily assess knowledge mastered by medical doctors, neglecting other essential professions. To address this gap, we introduce the Examinations for Medical PErsonnel in Chinese (EMPEC), a comprehensive healthcare knowledge benchmark featuring 157,803 exam questions across 124 subjects and 20 healthcare professions, including underrepresented roles like Optometrists and Audiologists. Each question is tagged for release time and source authenticity. We evaluated 17 LLMs, including proprietary and open-source models, finding that while models like GPT-4 achieved over 75% accuracy, they struggled with specialised fields and alternative medicine. Notably, we find that most medical-specific LLMs underperform their general-purpose counterparts in EMPEC, and incorporating EMPEC's data in fine-tuning improves performance. In addition, we tested LLMs on questions released after the completion of their training to examine their ability in unseen queries. We also translated the test set into English and simplified Chinese and analyse the impact on different models. Our findings emphasise the need for broader benchmarks to assess LLM applicability in real-world healthcare, and we will provide the dataset and evaluation toolkit for future research. Our data and code are in https://github.com/zhehengluoK/eval_empec.

1 Introduction

Recent advancements in Large Language Models (LLMs) have demonstrated the potential of LLM-based Artificial Intelligence (AI) in providing accurate answers to questions about world knowledge. These advancements are reflected in a series of studies and models, including but not limited to GPT-4, Gemini, Mistral, and Llama series (OpenAI, 2023; Google, 2023; Jiang et al., 2023; Touvron et al., 2023). To benchmark the internal knowledge of LLMs, multiple datasets (Hendrycks et al., 2020; Clark et al., 2018; Lin et al., 2021) have been introduced, focusing on their ability to measure and respond with accurate information. More recently, there has been a significant push towards adapting LLMs for use in the biomedicine domain (Singhal et al., 2023; Chen et al., 2023a; Xie et al., 2024a), exploring the possibility of deploying these models in real healthcare scenarios. To assess the effectiveness of LLMs in healthcare, considerable effort (Cai et al., 2023; Wang et al., 2023b; Jin et al., 2021; Kasai et al., 2023; Pal et al., 2022) has been invested in benchmarking their capabilities. Among these efforts, several studies (Cai et al., 2023; Wang et al., 2023b) have incorporated long-form diagnostic questions into their benchmarks to more evaluate the models' capacities to function akin to real physicians. In another line of work, multiple-choice questions have emerged as a straightforward and objective means of evaluation (Jin et al., 2021; Kasai et al., 2023; Pal et al., 2022). These studies mostly utilize questions from medical licensing exams, research papers, and textbooks to assess LLMs' knowledge and suitability to serve in a physician-like role.

Existing medical benchmarks are limited in scope and authenticity In real-world healthcare environments, medical doctors, while integral, represent only a fraction of the entire healthcare system. The International Standard Classification of Occupations (ISCO) by the International Labour Organization categorises health professionals into two Sub-major Groups, with medical doctors listed alongside ten other minor occupational groups. These include Nursing and Midwifery Pro-

[*]The first two authors contribute equally.
♠Correspondence: xieq@whu.edu.cn;

Findings of the Association for Computational Linguistics: ACL 2025, pages 9945–9958
July 27 - August 1, 2025

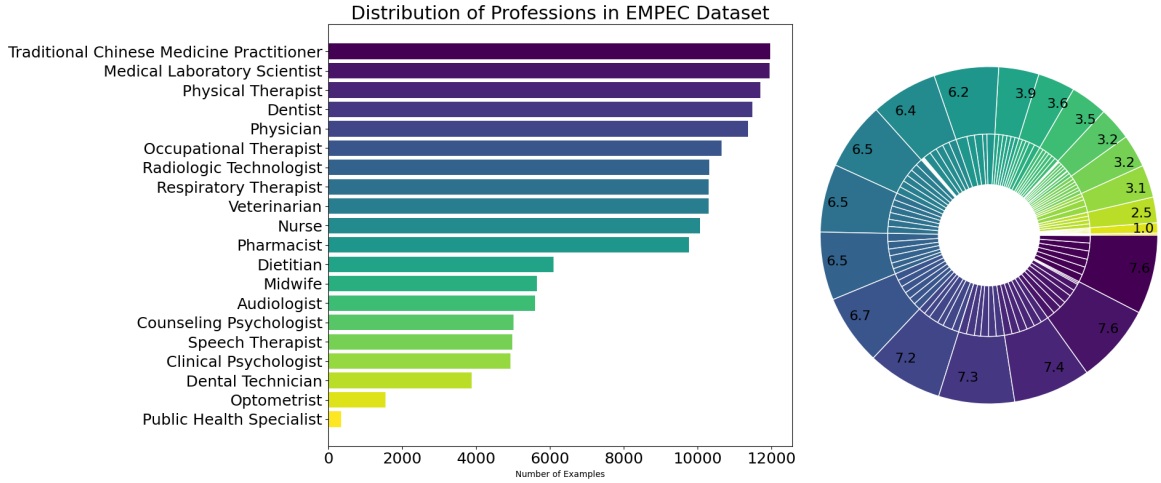

Figure 1: Distribution of professions in the EMPEC dataset. The left panel illustrates the total number of questions attributed to each healthcare profession within the dataset. The right panel provides a visual representation of the proportionate distribution of questions across the various professions.

fessionals and Paramedical Practitioners, among others. Each minor group comprises at least two unit groups, which in turn consist of multiple professions. This diversity in healthcare roles highlights that the expertise of medical doctors is just a subset of the broader healthcare knowledge spectrum. Nevertheless, as shown in Table 1, existing medicine-related benchmarks face challenges in diversity. For instance, MedQA (Jin et al., 2021) and MedBench (Cai et al., 2023) contain questions solely for physicians. MedMCQA (Pal et al., 2022) gathers questions for postgraduate medical students. CMB (Wang et al., 2023b) and CMExam (Liu et al., 2024) have made limited attempts to expand coverage beyond physicians to a maximum of five professions, leaving many health professionals unrepresented. Consequently, there remains a significant gap in our ability to assess the performance of LLMs across the broader spectrum of medical contexts. As a result, there remains a significant gap in our ability to assess the performance of LLMs across the broader spectrum of medical contexts. To more comprehensively gauge the effectiveness and applicability of LLMs in healthcare, it is crucial to expand the scope of benchmarks to include a wider range of professions within the healthcare system. Furthermore, despite being collected from credible sources, Table 1 reveals that most existing works cannot reference an authoritative source for their data, casting doubt on the authenticity of the collected problems. Additionally, essential metadata, such as release dates, are often missing. Therefore, the results from existing benchmarks fail to accurately reflect LLMs' performance across the entire healthcare system.

In response to the identified research gap, we introduce EMPEC, a comprehensive large-scale healthcare knowledge benchmark. EMPEC comprises 157,803 exam questions spanning 124 subjects across **20** healthcare professions. This comprehensive benchmark goes beyond the commonly assessed professions such as Physicians and Nurses, to include occupations like Optometrist and Audiologist often overlooked in previous assessments, thereby filling a critical void in existing benchmarks. The benchmark is originally in traditional Chinese; to broaden the applicability of our work beyond Chinese-speaking regions, we add an English translation of the test set generated by GPT-4. EMPEC stands out not only for its substantial size and extensive coverage but also for its authoritative and time-sensitive features. Each question within EMPEC is tagged with its release time, ensuring that the benchmark continuously integrates the latest questions from the source as soon as they are made available. This feature effectively mitigates the risk of data contamination (Nori et al., 2023) during evaluation processes, thereby maintaining the benchmark's relevance and accuracy over time. Furthermore, the provenance of every question in EMPEC is documented, with each item linked to its original release. This verification ensures the benchmark's reliability and authenticity, positioning it as a critical tool for the evaluation of LLMs within the healthcare sector.

We conducted extensive experiments on 17

Table 1: Review of existing healthcare-related benchmarks, ✓ represents the dataset that has the feature and ✗ represents it does not. Source-verifiable means the dataset can be verified from its acclaimed source. Trad and simp are short for traditional and simplified respectively.

Dataset	Resources	Language	#Professions	#Question	Source-verifiable
CMExam	Gov Publication	Simp Chinese	5	68,119	✗
CMB	Open database	Simp Chinese	4	280,839	✗
MedQA	Text Books	English, Chinese	1	61,097	✗
MedMCQA	Open website&Books	English	-	193,155	✓
MedBench	Gov Publication	Simp Chinese	1	40,041	✗
EMPEC	Gov Publication	Trad Chinese	**20**	157,803	✓

LLMs, including two proprietary models and 14 open-source models, evaluating their zero-shot performance. Additionally, we include one fine-tuned model using the training data from the EMPEC dataset. Our analysis contrasted the performance of medical domain LLMs with general LLMs, as well as models primarily trained on Chinese versus English data. Furthermore, we evaluated the models on both the full test set and a subset comprising the most recent questions, which could be reliably assumed to be absent from the models' training data. Finally, we compared the models' performance on questions in traditional Chinese with their performance on simplified Chinese and English versions of the same questions. Our findings on the experimental results are as follows:

I) GPT-4 leads the evaluation by achieving more than 75% accuracy while open-source LLMs are catching up with the frontier. **II)** While leading LLMs perform well in frequently encountered professions, EMPEC shows that they struggle with more specialized fields knowledge such as Dentists and Optometrists, and alternative medicine like Traditional Chinese Medicine Practitioners. **III)** Previous LLMs trained for the medical domain unexpectedly show inferior performance compared to their general-purpose counterparts. Moreover, incorporating EMPEC data into training significantly boosts model performance. **IV)** The results on questions released later than the time cutoff of several models reveal consistency with the overall performance trends observed in the EMPEC test set, suggesting that the models' performance on the test set can be extrapolated to predict their effectiveness in addressing unseen healthcare-related queries. **V)** The conversion from traditional to simplified Chinese characters appears to have a negligible impact on the performance of the models. While on the translated English version, models primarily trained in English enjoy an improvement. Despite the change in language, the accuracy trends by profession remain consistent.

2 Related work

Healthcare Knowledge Benchmark Multiple endeavours have been made to evaluate the medical knowledge in LLMs to propose question-answering(QA) datasets. Long-form QA datasets such as MedicationQA (Abacha et al., 2019), BioASQ (Krithara et al., 2023) are usually drafted and annotated by domain experts to assess LLMs via comparing their free-format response to the human written answers. Medical exams have become an ideal source for collecting medical QA materials due to their endorsement by national institutions and under rigorous scrutiny. Several benchmarks, including MedQA (Jin et al., 2021), CMExam (Liu et al., 2024), IgakuQA (Kasai et al., 2023), and Polish MFE (Rosoł et al., 2023) leverage medical exams from different nations to test knowledge, focussing primarily on physicians for the respective languages. MedBench (Cai et al., 2023) went beyond Chinese MLE to advanced physician exams such as the Resident Standardisation Training Examination, the Doctor In-Charge Qualification Examination, and real-world clinic cases encompassing examinations. CMB (Wang et al., 2023b) expands the assessment of medical knowledge by introducing exams for nurses, technicians, and pharmacists. Though of high quality, these existing benchmarks mostly focus on assessing a small proportion of health professions represented by physicians, leaving the knowledge mastered by other healthcare personnel unchecked, limiting these datasets' capability to comprehensively assess LLMs' knowledge in the entire healthcare system.

Knowledge-related Benchmarks for LLM Given the advance of LLMs, there was a multitude

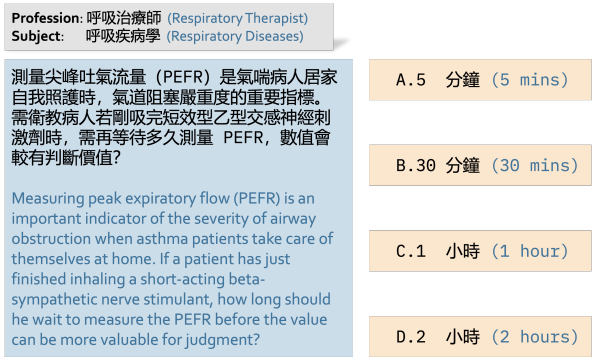

Figure 2: An example of questions in EMPEC, texts in blue are English translations of the original Chinese question and answers.

of works focussing on benchmarking knowledge within the models. MMLU (Hendrycks et al., 2020) and ARC (Clark et al., 2018) collected various questions from a diverse set of subjects aiming to comprehensively assess the world knowledge of models. GSM8K (Cobbe et al., 2021) collected grade school math word problems to benchmark models' mathematics reasoning ability. FinBen (Xie et al., 2024b) aggregated 35 datasets across 23 financial tasks aiming to fully assess LLMs' knowledge and capacity in finance. LawBench (Fei et al., 2023) compiled datasets to evaluate LLM's legal capacities from three cognitive levels. However, in healthcare, most existing benchmarks concentrate on examining physicians' expertise, being unable to assess models' knowledge in the whole healthcare system.

3 The EMPEC Dataset

Data Collection and Pre-processing EMPEC compiles officially published past exams for healthcare professionals in the Professional and Technical Examinations of Taiwan, Republic of China[1]. Each profession's exams consist of multiple related subjects designed to comprehensively assess the candidates medical knowledge and clinical skills for the profession. Tests for some professions are conducted biannually, and we have gathered exams held from 2011 to 2024 of issuance to streamline training and evaluation processes. To ensure the quality of EMPEC, we conducted the following pre-processing steps: 1) Excluding exams for questions requiring non-textual information such as images or tables; 2) For questions that shared the same premise, we added the premise for the all following questions. 3) We remove questions belonging to subjects like "Traditional Chinese literature" and "Pharmaceutical Administration and Regulations" to rule out problems that are irrelevant or only applied in the local region to calibrate EMPEC's concentration in healthcare knowledge. 4) We use Data-Juicer (Chen et al., 2024) to deduplicate questions that are highly similar to each other. Detailed process configuration is given in Appendix E. As a result, we have curated a collection of 157,803 multiple choice questions. In addition, to investigate whether the model's performance would be affected by language, we employed GPT-4 [2] to translate the test set to English using a simple prompt as "Translate the following question into English". We set the temperature at 0 to ensure reproducibility and asked domain experts to carefully review a subset of the English translations to assess their quality (see C for further details on the evaluation process). We also adopted *zhconv*[3] to convert the questions into simplified Chinese. This extensive collection covers 20 medical professions across 124 subjects. An example of questions in EMPEC is shown in Figure 2. Lastly, we will follow this pipeline to collect the latest questions issued from the officials.

Dataset Statistics In EMPEC, each question presents four options, with only one being the correct choice. Figure 1 illustrates the distribution of EMPEC questions in various professions. There are 11 professions taking more than 6% per cent, 7 professions taking about 3%. The questions for Optometrists and Public Health Specialists take less than 1% since the exams for the two professions were recently introduced into the national exams. In Appendix D, we explicitly show the subjects examined and the number of questions in each profession. The average subject of examination for each profession is more than 6, showing the diverse assessment of the expertise of each profession. Thus, EMPEC offers a more comprehensive coverage compared to prior studies (Wang et al., 2023b; Jin et al., 2021; Cai et al., 2023) which primarily focused on physicians. We split the dataset into 3 subsets, train, validation, and test. We first split the test set from the full dataset via stratified sampling by profession to ensure the same distribution. Then, we sampled 10 questions per profession

[1] https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx

[2] https://platform.openai.com/docs/models/gpt-4o

[3] https://github.com/gumblex/zhconv

from the rest data to form the validation set to facilitate a few-shot evaluation. The remaining data serves as the training subset, and the statistics are shown in Table. 2.

Table 2: Statistics of each split of EMPEC.

Split	#Profession	#Subject	#Year	#Question
Train	20	124	14	149,603
Validation	20	100	14	200
Test	20	124	14	8,000

Dataset Characteristics EMPEC has several advantages over existing medical QA benchmarks: 1) Comprehensive Healthcare Knowledge: Unlike MedBench (Cai et al., 2023) and CMExam (Liu et al., 2024), which mainly focus on physicians, EMPEC evaluates 20 healthcare professions, offering broader coverage. Existing benchmarks cover fewer than five occupations. 2) Extensive Question Pool: EMPEC includes over 157K questions from 124 subjects, surpassing CMExam, MedQA, and MedBench. 3) High-Quality Assurance: The questions are crafted by experts and vetted by Taiwan's Ministry of Examination. 4) Timestamp and Self-Growth: EMPEC tracks question release years and integrates new questions, allowing for assessment of data contamination in LLM training.

4 Benchmark

4.1 Tested Models

General LLMs We select leading proprietary LLMs GPT-4-turbo and GPT-3.5-turbo aiming to set an upper bound for LLMs' performance on EMPEC. For open-source models, We choose leading English-majored LLMs including Llama-3 (AI@Meta, 2024), Mistral (Jiang et al., 2023) as well as five Chinese-majored LLMs Yi (Young et al., 2024), Qwen1.5 (Bai et al., 2023), Baichuan2 (Yang et al., 2023), InternLM2 (Cai et al., 2024), and Ziya-Llama(IDEA-CCNL, 2021). Both Yi and Qwen1.5 have shown performance on par with or better than GPT-4 on multiple Chinese benchmarks.

Medical Domain LLMs As further training LLMs on in-domain data usually brings improvement in domain-specific tasks (Wang et al., 2023a; Chen et al., 2023b), we further examined the LLMs that have been trained or fine-tuned on biomedical texts. BioMistral-7B (Labrak et al., 2024) is based on Mistral 7B Instruct v0.1 and further pre-trained on PubMed Central data. HuatuoGPT2 (Chen et al., 2023a). The 13B and 34B versions of HuatuoGPT2 are pre-trained on Baichuan2-13B and Yi-34B respectively using more than 5 million synthetic medical instruction tuning data. HuatuoGPT2-13B is reported to outperform GPT3.5 by more than 20% on two Chinese medical benchmarks CMB (Wang et al., 2023b) and CMExam (Liu et al., 2024). MMedLM2 (Qiu et al., 2024), based on InternLM2, is further pre-trained on 22.5 billion tokens of health-related text across 6 languages including Chinese. MedGPT (Xu, 2023) is based on Ziya-Llama-13B (IDEA-CCNL, 2021), which is a Llama variant with an expanded vocabulary and further pre-trained on Chinese and English texts. MedGPT, on top of Ziya-Llama-13B, is fine-tuned on 240 million Chinese and English medical instruction-tuning data. Qwen1.5-7B-SFT, we fine-tuned a Qwen1.5-7B using the training data of EMPEC.

Specifically, we contrast the medical LLMs with their general counterparts like Baichuan2 and Yi for HuatuoGPT2, ZiyaLlama for MedGPT, InternLM2 for MMedLM2, Qwen1.5-Chat for Qwen1.5-SFT to explore the effect of the domain-adaption training of these models.

4.2 Evaluation settings

We use the *0125* version of GPT-turbo models. To facilitate zero-shot prompting, we choose the "Chat" or "Instruct" versions of Llama-3, Mistral(v0.2), Baichuan2, Qwen1.5, and Yi. We fine-tuned Qwen1.5-7B on the EMPEC training subset for 3 epochs using the learning rate of 1e-4 on 2 A100 GPUs. We use vLLM 0.4.1 (Kwon et al., 2023) and enable greedy decoding to ensure the stability and reproducibility of the results. The prompt used in zero-shot prompting and fine-tuning is in Appendix A. The evaluation procedure follows common practice in previous work (Bai et al., 2023; Huang et al., 2023) to extract the correct option from the model's response. We allow the model to generate up to 256 tokens for each question. Moreover, as existing benchmarks suffer the problem of data leakage (Nori et al., 2023; Zhang et al., 2024), we explicitly conduct the evaluation of the question from the exams held in 2024 on models of which the pre-training data cut-off is before 2024. During our evaluation, questions that are filtered by the API provider or models fail to respond with the correct choice are deemed incorrect. A baseline is set by randomly choosing a choice from A to D. We did not examine few-shot as most

Table 3: Accuracy of each model on the test set of EMPEC, split by profession. C stands for Chat version and I stands for Instruction version. Bch2 is short for Baichuan2.

Split	Accuracy∗																	
	GPT4	GPT3.5	Qwen1.5			Llama3-I		Yi-C		HuatuoGPT2		Bch2-C	Mistral-I	InternLM2	MMedLM2	MedGPT	Ziya	BioMistral
	-	-	70B-C	7B-C	7B-SFT	70B	8B	34B	6B	34B	13B	13B	7B	7B	7B	7B	7B	7B
Nurse	**82.75**	63.53	77.65	61.18	69.22	70.20	52.75	73.53	48.04	41.96	17.25	55.10	40.20	62.16	58.43	24.31	29.41	18.82
Dentist	**69.27**	49.83	53.30	47.22	53.99	57.47	49.31	56.60	34.38	33.85	18.40	42.71	35.24	46.18	47.22	24.65	23.61	18.75
Midwife	**85.31**	58.04	69.23	49.65	62.24	72.73	53.85	69.93	39.16	41.61	17.48	49.30	36.71	55.24	51.05	23.08	23.43	16.08
Physician	**89.90**	67.42	72.30	58.01	68.47	83.62	68.29	73.87	46.34	42.86	14.81	54.70	45.12	62.89	62.37	25.09	25.78	20.91
Dietitian	**74.76**	53.40	70.55	52.75	61.49	65.37	47.90	63.43	39.16	44.66	20.71	50.49	37.54	56.96	52.43	29.77	20.71	18.77
Pharmacist	**75.86**	58.15	60.36	49.09	57.14	72.03	53.32	64.59	42.05	45.88	15.90	45.07	39.44	56.74	51.71	30.38	26.76	21.73
Audiologist	**72.18**	53.17	63.03	44.37	54.23	59.51	44.72	57.39	34.51	34.15	19.37	40.14	33.45	49.30	44.37	23.59	20.77	13.03
Optometrist	**64.10**	51.28	51.28	44.87	53.85	52.56	46.15	48.72	37.18	39.74	16.67	41.03	25.64	44.87	42.31	20.51	21.79	14.10
Veterinarian	**78.35**	55.56	63.22	49.81	63.41	63.22	50.38	60.73	38.70	36.02	17.62	43.87	38.12	53.07	52.49	21.26	25.29	21.07
Speech Therap	**69.05**	46.43	54.76	40.87	59.13	52.38	36.90	51.19	23.81	33.33	19.05	38.49	29.37	40.87	41.67	27.78	22.22	15.87
Dental Tech	**59.39**	41.12	58.88	50.25	53.81	46.70	44.16	52.28	36.55	35.03	13.71	38.07	23.86	44.67	43.65	31.47	26.90	17.77
Phys Therap	**70.25**	51.43	52.61	52.27	58.49	60.67	52.10	59.33	37.98	36.13	19.50	46.89	36.47	53.28	50.76	28.24	24.54	19.66
Resp Therap	**75.43**	53.74	52.59	48.56	57.01	61.61	49.52	56.62	36.85	35.32	20.54	42.42	39.16	49.90	47.22	24.38	24.38	17.66
Clin Psych	**87.65**	67.73	77.29	59.76	65.74	72.11	59.36	72.11	46.22	43.03	19.92	54.98	45.82	67.33	64.54	31.47	31.87	23.11
Occup Therap	**76.85**	58.89	60.93	53.52	65.56	62.41	51.85	61.85	39.44	36.48	20.56	50.37	40.56	56.11	54.63	30.37	26.48	19.81
Rad Tech	**76.97**	54.32	58.73	43.57	59.12	64.68	51.44	54.70	30.33	33.40	17.66	38.96	34.36	51.06	48.37	25.53	20.92	15.55
Counsel Psych	**79.30**	60.94	75.78	61.72	68.75	67.19	60.55	73.44	53.52	42.19	16.02	60.55	43.75	64.84	61.72	31.25	28.91	16.41
PH Spec	**82.35**	70.59	58.82	52.94	76.47	52.94	58.82	64.71	47.06	41.18	41.18	52.94	47.06	52.94	41.18	41.18	17.65	35.29
Med Lab Sci.	**85.17**	67.71	68.86	55.68	63.26	76.77	58.81	66.72	40.53	43.00	18.12	54.04	44.81	60.30	58.32	30.31	26.19	22.57
TCM Prac	50.08	36.24	**63.26**	39.37	54.53	44.81	34.27	55.85	32.13	21.91	13.84	43.16	24.71	49.59	44.98	16.31	17.30	8.40
Micro Avg.	**75.35**	55.66	63.24	50.79	60.84	64.46	51.41	62.29	38.79	37.45	17.81	47.20	37.44	54.50	52.08	26.08	24.51	18.25
Marco Avg.	**75.25**	55.96	63.17	50.77	61.27	63.02	51.23	61.90	39.18	38.06	18.86	47.17	37.09	53.94	51.02	27.00	24.26	18.74

∗ A random guess baseline gets 24.96% micro average accuracy

of the tested models are specifically fine-tuned for following instructions.

5 Analysis

In general, all the tested models show unsatisfying performance. The average accuracy of 12 out of the 17 tested models is under 60% on EMPEC where random guessing can reach about 25%. The results suggest the difficulty of our benchmark and the healthcare knowledge gap in existing LLMs.

5.1 Results of general LLMs

From the accuracy of each profession and the average overall profession, GPT-4 shows an evident lead among the tested models, with only a second Qwen1.5-70B-Chat in Traditional Chinese Medicine Practitioner. Another proprietary model, GPT-3.5 shows a nearly 20% performance drop from GPT 4. Open-source models are far behind GPT-4, the accuracy of the best-performing models, Qwen1.5-70B-Chat and Llama3-70B-Instruct, is around 63% — 12% less than GPT-4. However, gained from training on a large proportion of Chinese data, Yi-34B and Qwen1.5-70B beat GPT 3.5. Observing the results of Yi, Qwen1.5, and Llama-3, we see that increasing the model size can benefit models' performance as expected. Among models of around 7B size, InternLM2, Qwen1.5-Chat, and Llama3-8B-Instruct achieve over 50% accuracy while Mistral and Yi are slightly lower than 40%.

5.2 Results of medical domain LLMs

HuatuoGPT2, MMedLM2, and MedGPT, despite being based on Chinese-focused LLMs and specifically trained on medical data, show poor performance in EMPEC. The results of MedGPT are slightly higher than random guess while HuatuoGPT2-13B even underperforms the random baseline. MMedLM2 performs best among medical LLMs, achieves 45% accuracy with 7B parameters, but still lags behind Yi-6B and Qwen1.5-7B. We noticed that part of the reason for poor performance is the loss of instruction-following ability. For instance, HuatuoGPT2, despite being able to generate plausible response, sometimes deviates from the questions and often fails to conclude an answer. HuatuoGPT2 underperforms Baichuan-2-13B-Chat and Yi-34B-Chat which shares the same base model but fine-tuning on general instruction data. The findings are different from the results reported in Wang et al. (2023b) where HuatuoGPT2 outperforms Baichuan2 by a large margin in Chinese medical questions. In addition, InternLM2 leads MMedLM2 by more than 2% accuracy while fine-tuned MMedLM2 outperforms InternLM2 on MedQA (Jin et al., 2021). Only MedicalGPT outperforms Ziya-Llama by around 2%. In conclusion, we do not observe evident enhancement on EMPEC brought by fine-tuning or pre-training in the three works, contrasting the findings on existing benchmarks like CMB (Wang et al., 2023b). Therefore, we argue that, compared to the existing Chinese healthcare benchmark, our new EMPEC provides a more robust platform for the eval-

uation of domain-adaptation LLMs. This finding further suggests that the current improvement in domain adaptation could result in an over-fit of the tested distribution, which might not hold once the distribution switches. Moreover, our fine-tuned Qwen1.5-7B model achieves 61% accuracy, nearly 11% better than the Chat counterpart and close to Qwen1.5-70B-Chat, suggesting that the training data of EMPEC can be an effective supplement of current medical domain adaptation endeavours.

5.3 Results by profession

We further examined the performance of the models by profession. While GPT-4 prevails on 19 of the 20 professions, Qwen1.5-70B-Chat achieves the highest accuracy in Traditional Chinese Medicine Practitioner. Then we rank the accuracy of each profession within every model and then aggregate rankings across models. The best three performed professions are Clinical Psychologist, Physician, and Nurse while the models mostly struggle to answer questions for Dental technician, Speech Therapist, Optometrist, Dentist, and TCM Practitioner, which seem connect more to specific medical subject. We assume this difference is rooted in the frequency of knowledge in the training data of these models and the rareness of knowledge of these bad-performing professions as Kandpal et al. (2023) has suggested. Take GPT-4 as an example, it achieves nearly 90% accuracy for physicians of which related-documents are rich on the Internet and books, while its performance on rarely touched professions like Dental technician and Traditional Chinese Medicine Practitioner is lower than 60%. Interestingly, all tested medical domain LLMs except Qwen1.5-7B-SFT do not show a better performance compared to their general counterparts even on physicians which is the focus of their finetuning data. However they generally unsatisfying performance in each profession indicating the importance of data diversity in building an LLM for healthcare. By breaking the overall results into professions, EMPEC is able to precisely detect LLM's knowledge gap in the healthcare domain.

5.4 Results on the unseen questions

During the collection of EMPEC, we found that some of the past questions are freely available on the Internet, the possibility of the data having been seen by LLMs cannot be ruled out. Thus, we explicitly composed a dataset containing 3497 questions issued in the exams in 2024, which should not be seen by the models of which cut-offs are before 2024. The statistics of this data set are given in Table 6. We tested GPT-4-turbo(cutoff on Dec. 2023), GPT-3.5-turbo(cutoff on Sept. 2021), Yi (cutoff on Dec. 2023), Baichuan-2 (released on Sept. 2023), and Mistral (released at Dec. 2023) on the questions released in 2024. It is important to note that this split might have overlapped with the test set of EMPEC and the included professions are limited, as the exams for the rest professions have not been held yet. The results are in Table 4. In general, we observed that the results of questions 2024 are very close to the results obtained from the test set across assessed models. In other words, the performance rankings are held as in the test set. Moreover, the professions where models perform well, such as Physician and Nurse, perform poorly, such as Dentist and Traditional Chinese Medicine Practitioner also keep in the latest questions set. The findings strongly suggest the robustness of the EMPEC test set, as the performance of the models is consistent with that of the unseen data. In addition, it also indicates that the tested models do not gain their performance by reciting seen examples.

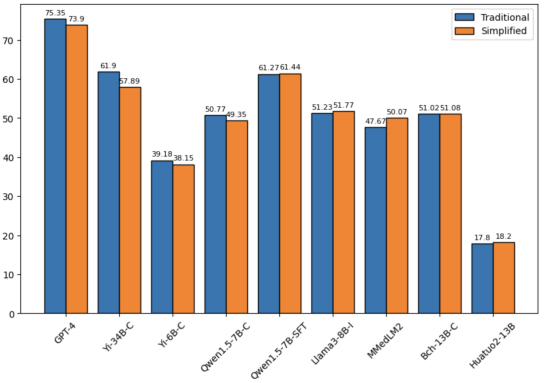

Figure 3: Micro average accuracy of models on traditional Chinese and simplified Chinese. C stands for Chat version and I stands for Instruction version

5.5 Results by language

As many of the tested models are trained primarily in English, we tested them to exmaine the impact brought by transfering to English. Figure 4 shows the performance of models in the English-translated test set, more results are shown in Table 5, we observed that GPT-4, Llama 3, Yi, MMedLM 2, and Mistral demonstrated improved performance following translation. However, InternLM 2 and Qwen 1.5-Chat exhibited a decline. It

Table 4: Performance of models on questions from 2024. Colored numbers are the results of questions from exams in 2024 while the black ones are the results of the test set. Red indicates higher while blue suggests degrade.

Profession	Accuracy				
	GPT 4	GPT 3.5	Yi-Chat-34B	Baichuan2-Chat-13B	Mistral-Instruct-7B-v0.2
Phys Therap	73.05 (70.25)	50.78 (51.43)	54.34 (59.33)	49.00 (46.89)	36.97 (36.47)
Rad Tech	72.58 (76.97)	51.52 (54.32)	58.68 (54.70)	44.63 (38.96)	34.44 (34.36)
Physician	85.74 (89.90)	63.39 (67.42)	72.18 (73.87)	48.74 (54.70)	44.77 (45.12)
TCM Prac	57.05 (50.08)	36.58 (36.24)	50.65 (55.85)	46.97 (43.16)	24.03 (24.71)
Dentist	67.22 (69.27)	47.29 (49.83)	50.82 (56.60)	39.76 (42.71)	38.82 (35.24)
Pharmacist	76.24 (75.86)	58.01 (58.15)	63.54 (64.59)	47.51 (45.07)	38.40 (39.44)
Med Lab Sci.	85.97 (85.17)	63.92 (67.71)	67.93 (66.72)	47.88 (54.04)	43.21 (44.81)
Dietitian	75.93 (74.76)	55.51 (53.40)	60.29 (63.43)	45.59 (50.49)	40.44 (37.54)
Nurse	87.76 (82.75)	71.31 (63.53)	78.48 (73.53)	56.54 (55.10)	43.46 (40.20)
Micro Avg.	75.12 (74.54)	54.48 (55.47)	61.11 (62.82)	47.07 (47.70)	37.95 (37.44)

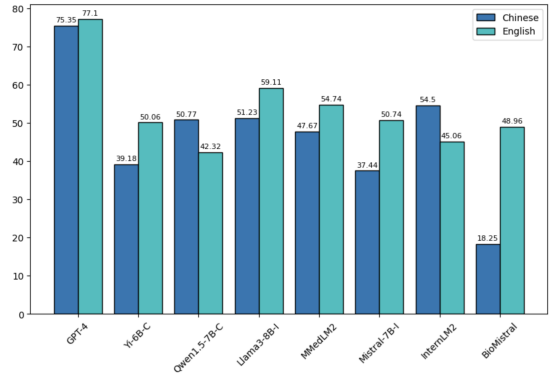

Figure 4: Micro average accuracy of models on traditional Chinese and English. C stands for Chat version and I stands for Instruction version

is evident that models primarily trained in English or continual-pretrained in English like MMedLM2 benefit from this transition. However, it is crucial to note that, in English, accuracy trends by profession remain consistent. Models continue to struggle with specialities such as Dietitian, Audiologist, Optometrist, and Veterinarian, further underscoring the unique challenges posed by the EMPEC dataset. Furthermore, the slightly lower performance of BioMistral compared to Mistral reinforces our assertion that current medical LLMs do not consistently outperform their general-purpose counterparts in comprehensive healthcare knowledge.

The difference between traditional Chinese and simplified Chinese can affect the readers' analytic skills (Liu and wen Hsiao, 2012). We further investigated the performance of a subset of tested models on the test set in traditional Chinese and simplified Chinese. The results are shown in Figure 3. From traditional to simplified, contrary effects are observed among the models with GPT-4, Yi-Chat, and Qwen1.5-Chat suffering slight performance drops while HuatuoGPT2, MMedLM2, Llama3-Instruct, Qwen1.5-SFT enjoying slight improvement. However, the performance across these models between the two kinds of Chinese does not vary greatly though the traditional and simplified characters do not share the same tokens. We assume that the character difference in traditional and simplified Chinese does not affect the model's performance as the co-occurrence relations of these characters are similar in the two kinds of Chinese.

6 Conclusion and Discussions

We introduced EMPEC, the most comprehensive healthcare knowledge benchmark to date, encompassing 157,803 questions across 124 subjects and 20 healthcare occupations. EMPEC goes beyond the physician-centric focus of previous medical benchmarks to provide a holistic assessment of the knowledge of LLM in a wide spectrum of health disciplines. Our extensive experiments, which tested both proprietary and open-source models, revealed several important findings. Although the best general LLMs performed reasonably well in common professions such as Physicians and Nurses, they struggled with rarer specialities like Dental Technicians, Optometrists, and TCM Practitioners. This highlights the need for more work to improve the knowledge of LLMs in healthcare beyond the expertise of physicians that has been the primary focus so far. Interestingly, we found that existing medical domain-specific LLMs did not perform better on EMPEC compared to their general counterparts not specialised in healthcare. Furthermore, our experiments indicate that the models' performance on the EMPEC test set can predict their effectiveness in addressing unseen healthcare-related queries. We also found that the transition from traditional to simplified Chinese characters

	InterLM2-7B-Chat	BioMistral-7B	MMedLM2-7B	Yi-Chat-6B	Qwen1.5-7B-Chat	Mistral-7B-Instruct v0.2	GPT4	Llama3-8B-Instruct	GPT3.5
Nurse	57.45	56.86	63.33	60.20	44.71	62.35	**82.55**	71.37	71.76
Dentist	40.28	45.49	45.49	42.36	38.54	46.70	**74.13**	53.65	53.30
Midwife	51.05	54.55	58.39	56.29	45.45	56.64	**84.27**	61.54	66.43
Physician	49.48	55.05	65.16	60.80	49.83	55.40	**90.77**	75.09	69.86
Dietitian	48.87	50.16	55.66	51.13	42.39	51.46	**77.35**	58.90	64.08
Pharmacist	46.98	48.79	54.44	50.00	41.73	53.63	**76.61**	57.46	60.89
Audiologist	42.05	44.17	50.18	42.40	34.28	42.05	**72.08**	49.47	56.89
Optometrist	34.62	38.46	43.59	50.00	53.85	52.56	**65.38**	51.28	60.26
Veterinarian	40.23	47.89	52.11	50.00	43.68	50.38	**81.23**	60.34	65.90
Speech Therap	40.87	38.10	45.24	36.51	30.16	40.08	**75.40**	50.79	54.37
Dental Tech	34.52	42.64	47.21	46.19	40.10	41.62	**68.02**	46.19	47.72
Phys Therap	41.32	50.76	54.97	46.37	39.63	45.70	**75.04**	57.17	57.50
Resp Therap	42.12	44.81	47.88	45.96	39.04	45.77	**74.81**	51.54	56.92
Clin Psych	54.40	63.60	68.80	58.80	54.00	64.00	**85.20**	72.00	76.00
Occup Therap	48.15	50.19	55.56	52.78	40.19	51.48	**78.70**	58.89	65.93
Rad Tech	44.51	51.25	54.72	47.98	39.50	47.21	**75.72**	55.49	57.23
Counsel Psych	54.69	57.81	62.11	59.38	43.75	56.25	**85.55**	67.19	71.09
PH Spec	70.59	52.94	70.59	52.94	52.94	64.71	64.71	**82.35**	**82.35**
Med Lab Sci.	50.74	57.33	63.10	58.32	45.80	56.84	**86.00**	69.03	71.00
TCM Prac	30.31	28.34	36.24	32.78	26.85	29.98	**51.57**	32.45	34.76
Marco Avg.	46.16	48.96	54.74	50.06	42.32	50.74	**76.25**	59.11	62.21
Micro Avg.	45.06	48.96	54.15	49.79	41.08	49.70	**77.10**	58.26	60.87

Table 5: Performance of each model on the English translation of the test set of EMPEC, split by profession.

had negligible impact on model performance. EMPEC lays the foundation for the development of advanced LLMs for the healthcare domain, providing a more comprehensive and nuanced evaluation of their capabilities in a wide range of health professions.

Limitations EMPEC has several limitations: Some professions included in EMPEC have a relatively smaller representation compared to others, which may restrict EMPEC's effectiveness in evaluating knowledge in those specific professions. The English translation is generated by LLMs without further examination of experts.

Societal Impacts Although EMPEC is designed to improve the evaluation of LLM in the medical field, it should not be used to assess individual medical competence or to diagnose any patient. Any conclusions drawn from models trained on this dataset should take into account its limitations. The dataset's use should be confined to research purposes to prevent potential misuse. In addition, as the simplified Chinese version and English version are obtained including conversion and translation, we cannot guarantee the accuracy are fully aligned with the original questions.

The datasets are built using publicly available open-source data from Taiwan's Ministry of Examination. Their use is encouraged, provided the source is clearly cited. Furthermore, the questions in the datasets do not contain any private information.

References

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*, pages 25–29.

AI@Meta. 2024. Llama 3 model card.

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*.

Yan Cai, Linlin Wang, Ye Wang, Gerard de Melo, Ya Zhang, Yanfeng Wang, and Liang He. 2023. Medbench: A large-scale chinese benchmark for evaluating medical large language models. *ArXiv*, abs/2312.12806.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. 2024. Data-juicer: A one-stop data processing system for large language models. In *International Conference on Management of Data*.

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. *Preprint*, arXiv:2311.09774.

Zeming Chen, Alejandro Hern'andez Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Kopf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models. *ArXiv*, abs/2311.16079.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Zhiwei Fei, Xiaoyu Shen, D. Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *ArXiv*, abs/2309.16289.

Gemini Team Google. 2023. Gemini: A family of highly capable multimodal models. *ArXiv*, abs/2312.11805.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Fanchao Qi, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *ArXiv*, abs/2305.08322.

IDEA-CCNL. 2021. Fengshenbang-lm. `https://github.com/IDEA-CCNL/Fengshenbang-LM`.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. 2023. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *arXiv preprint arXiv:2303.18027*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. Bioasq-qa: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. 2024. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36.

Tianyin Liu and Janet Hui wen Hsiao. 2012. The perception of simplified and traditional chinese characters in the eye of simplified and traditional chinese readers. *Journal of Vision*, 12:533–533.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375.

OpenAI. 2023. Gpt-4 technical report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *arXiv preprint arXiv:2402.13963*.

Maciej Rosoł, Jakub S Gąsior, Jonasz Łaba, Kacper Korzeniewski, and Marcel Młyńczak. 2023. Evaluation of the performance of gpt-3.5 and gpt-4 on the polish medical final examination. *Scientific Reports*, 13(1):20512.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Hao Wang, Chi-Liang Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023a. Huatuo: Tuning llama model with chinese medical knowledge. *ArXiv*, abs/2304.06975.

Xidong Wang, Guimin Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Cmb: A comprehensive medical benchmark in chinese. *ArXiv*, abs/2308.08833.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. 2024a. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Zi-Zhou Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, Zhiyang Deng, Yuechen Jiang, Zhiyuan Yao, Haohang Li, Yangyang Yu, Gang Hu, Jiajia Huang, Xiao-Yang Liu, Alejandro Lopez-Lira, Benyou Wang, Yanzhao Lai, Hao Wang, Min Peng, Sophia Ananiadou, and Jimin Huang. 2024b. The finben: An holistic financial benchmark for large language models. *ArXiv*, abs/2402.12659.

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

Ai Ming Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Hai Zhao, Hang Xu, Hao-Lun Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kuncheng Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Pei Guo, Ruiyang Sun, Zhang Tao, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yan-Bin Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *ArXiv*, abs/2309.10305.

01.AI Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.

A Prompt

The prompt used in the zero-shot evaluation and supervised fine-tuning is shown in Fig. 5.

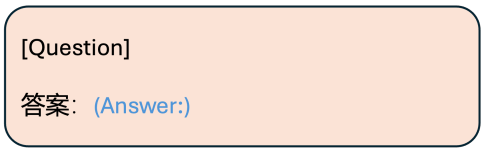

Figure 5: The prompt used in the zero-shot evaluation and supervised fine-tuning. The texts in blue are the English translations of the Chinese content.

B Statistics of EMPEC Questions Issued in 2024

The detailed statistics of EMPEC

Professors	Number of questions
Physical Therapist	449
Radiologic Technologist	363
Physician	478
Traditional Chinese Medicine Practitioner	462
Dentist	425
Pharmacist	362
Medical Laboratory Scientist	449
Dietitian	272
Nurse	237

Table 6: Number of questions for each profession in questions issued in 2024

C Human assessment of English translation

We conducted a manual evaluation of the English translations generated by GPT-4 using 10%(800) of the EMPEC test set. The evaluation was performed by a PhD-level medical student who is a native Chinese speaker and fluent in professional biomedical English. No errors were identified in the translations reviewed.

D Subject distributions in EMPEC

We show the distribution of subjects in EMPEC in Table 7.

Table 7: Subjects examined for each profession

Profession	Subjects	#Subjects	#Questions
Traditional Chinese Medicine Practitioner	Basic Chinese Medicine (1-2); Clinical Chinese Medicine (1-4); Pharmacy and Biopharmaceutics	7	11954
Medical Laboratory Scientist	Clinical Hematology and Blood Bank; Biochemistry and Clinical Biochemistry; Microbiology and Clinical Microbiology; Clinical Serum Immunology and Clinical; Medical Molecular Testing and Clinical; Clinical Physiology and Pathology; Clinical Mirror Examination	7	11938
Physical Therapist	Basic Physical Therapy; Cardiopulmonary and Pediatric Disease Therapy; Orthopedic Disease Physical Therapy; Neurological Disease Physical Therapy; Introduction to Physical Therapy; Physical Therapy Techniques	6	11698
Dentist	Dentistry (1-6)	6	11468
Physician	Medicine(1-6); Clinical Psychology Special Topics	7	11366
Occupational Therapist	Anatomy and Physiology; Psychological Disability Occupational Therapy; Occupational Therapy Techniques; Introduction to Occupational Therapy; Pediatric Occupational Therapy; Physiological Disability Occupational Therapy	6	10646
Radiologic Technologist	Basic Medical Science; Radiation Therapy Principles and Techniques; Nuclear Medicine Diagnosis Principles and Techniques; Medical Physics and Radiation Safety; Radiological Equipment; Radiological Diagnosis Principles and Techniques	6	10307
Respiratory Therapist	Cardiopulmonary Basic Medical Science; Respiratory Principles and Applications; Intensive Respiratory Therapy; Respiratory Therapy Equipment; Respiratory Diseases; Basic Respiratory Therapy	6	10301
Veterinarian	Veterinary Laboratory Diagnosis; Veterinary Pharmacology; Veterinary General Diseases; Veterinary Infectious Diseases; Veterinary Pathology; Veterinary Public Health	6	10292
Nurse	Internal and Surgical Nursing; Basic Nursing; Basic Medical Science; Mental Health and Community Health Nursing; Obstetric and Pediatric Nursing; Overview of Basic Medical Science; Obstetrics, Psychiatry and Community; Overview of Basic Nursing; Overview of Internal and Surgical Nursing	9	10066
Pharmacist	Pharmacotherapy; Pharmacy; Dispensing and Clinical Pharmacy; Pharmacology and Pharmaceutical Chemistry; Pharmacy and Biopharmaceutics	5	9767
Dietitian	Group Meal Design and Management; Diet Therapy; Nutrition; Food Hygiene and Safety; Physiology and Biochemistry; Public Health Nutrition	6	6088
Midwife	Midwifery (1-2); Nursing for All Specialties; Basic Medical Science; Basic Nursing	5	5642
Audiologist	Basic Audiology; Hearing and Language Communication Disorders; Health of Auditory and Balance Systems; Electrophysiological Audiology; Behavioral Audiology; Principles and Practice of Hearing Aids	6	5600
Counseling Psychologist	Counseling and Psychotherapy Theories; Counseling and Psychotherapy Practice and; Human Behavior and Development; Group Counseling and Psychotherapy; Case Assessment and Psychological Evaluation; Psychological Foundations of Counseling; Mental Health and Abnormal Psychology; Mental Health; Psychological Testing and Assessment; Counseling and Psychotherapy Practice	10	5014
Speech Therapist	Articulation and Fluency Disorders; Basic Linguistics; Communication Disorders Overview; Neurological Communication Disorders; Child Language Disorders; Voice and Swallowing Disorders	6	4973
Clinical Psychologist	Special Topics in Clinical Psychology (1-2); Clinical Psychology Special Topics (1-2); Basic Clinical Psychology	5	4923
Dental Technician	Dental Technology (1-4)	4	3885
Optometrist	Optometry; Low Vision; Eye Anatomy, Physiology and Ethics; Contact Lens and Dispensing; Visual Optics	5	1538
Public Health Specialist	Health Administration and Management; Epidemiology; Environmental and Occupational Health; Biostatistics; Health Social Behavior	5	337
Total		124	157803

E Example of Deduplicate Configuration

```
# global parameters
project_name: 'all'                                     # project name for distinguish your configs
dataset_path: '/path/to/your/dataset'                   # path to your dataset directory or file
export_path: '/path/to/result/dataset.jsonl'            # path to processed result dataset.
np: 4                                                   # number of subprocess to process your dataset
open_tracer: true                                      # whether to open the tracer to trace the changes
                                                        # It might take more time when opening tracer


# process schedule: a list of several process operators with their arguments
process:
  # Mapper ops. Most of these ops need no arguments.
  - clean_email_mapper:                                 # remove emails from text.
  - clean_links_mapper:                                 # remove web links from text.
  - clean_copyright_mapper:                             # remove copyright comments.
  - expand_macro_mapper:                                # expand macro definitions in Latex text.
  - fix_unicode_mapper:                                 # fix unicode errors in text.

  # Filter ops
  - average_line_length_filter:                         # filter text with the average length of lines
      max_len: 1500
  - language_id_score_filter:                     # filter text in specific language with language scores
      lang: zh                                          # keep text in what language
      min_score: 0.95                                    # the min language scores to filter text
  - maximum_line_length_filter:                         # filter text with the maximum length of lines
      #min_len: 10                                      # the min length of filter range
      max_len: 7328                                     # the max length of filter range
  - special_characters_filter:                  # filter text with special-char ratio out of specific range
      #min_ratio: 0.0                                   # the min ratio of filter range
      max_ratio: 0.3                                    # the max ratio of filter range
  - text_length_filter:                                 # filter text with length out of specific range
      min_len: 10                                       # the min length of filter range
      max_len: 10000                                    # the max length of filter range
  - document_simhash_deduplicator:                   # deduplicate text samples using SimHash-LSH method
      tokenization: space                               # tokenization method for text.
      window_size: 6                                    # window size of shingling
      num_blocks: 6                                     # number of blocks in SimHash computing
      hamming_distance: 4                               # the max hamming distance to regard 2 samples
     ignore_pattern: null                          # whether to ignore sub-strings with specific pattern
```