Large Language Models in Healthcare: A Comprehensive Benchmark

Fenglin Liu¹, Hongjian Zhou¹, Yining Hua², Omid Rohanian¹, Lei Clifton³, David A. Clifton¹
¹Institute of Biomedical Engineering, University of Oxford, UK
² Harvard T.H. Chan School of Public Health, USA
³ Nuffield Department of Population Health, University of Oxford, UK

Abstract

The adoption of large language models (LLMs) to assist clinicians has attracted remarkable attention. Existing works mainly adopt the close-ended question-answering task with answer options for evaluation. However, in real clinical settings, many clinical decisions, such as treatment recommendations, involve answering open-ended questions without pre-set options. Meanwhile, existing studies mainly use accuracy to assess model performance. In this paper, we comprehensively benchmark diverse LLMs in healthcare, to clearly understand their strengths and weaknesses. Our benchmark contains seven tasks and thirteen datasets across medical language generation, understanding, and reasoning. We conduct a detailed evaluation of existing sixteen LLMs in healthcare under both zero-shot and few-shot (i.e., 1,3,5-shot) learning settings. We report the results on five metrics (i.e. matching, faithfulness, comprehensiveness, generalizability, and robustness) that are critical in achieving trust from clinical users. We further invite medical experts to conduct human evaluation.

1 Introduction

Large language models (LLMs), such as ChatGPT (Brown et al., 2020; OpenAI, 2023b), LLaMA (Touvron et al., 2023a), and PaLM (Chowdhery et al., 2022), are increasingly being recognized for their potential in healthcare to aid clinical decision-making and provide innovative solutions for complex healthcare problems (Patel et al., 2023; Shen et al., 2023), e.g., discharge summary generation (Patel and Lam, 2023), health education (Safranek et al., 2023), and care planning (Fleming et al., 2023). Several recent efforts have been made to fine-tune publicly available general LLMs, e.g., LLaMA (Touvron et al., 2023b) and ChatGLM (Tsinghua KEG, 2023), to develop medical LLMs (Li et al., 2023b), MedAlpaca (Han et al., 2023), BenTsao (Wang et al., 2023a), and ClinicalCamel (Toma et al., 2023). Previous research shows that medical LLMs outperform human experts across a variety of medical tasks. In particular, MedPrompt (Nori et al., 2023) and MedPaLM-2 (Singhal et al., 2023) have respectively achieved a competitive accuracy of 90.2 and 86.5 compared to human experts 87.0 (Wu et al., 2023) on the United States Medical Licensing Examination (USMLE) (Jin et al., 2021).

Admittedly, responsibility and reliability are essential requirements for tools designed to assist clinicians. Despite the promising results of existing medical LLMs, several issues need to be addressed for the responsible and reliable use of LLMs in assisting clinicians:

• (i) Limited evaluation: Most existing works only focus on evaluating LLM performance in the close-ended medical question answering (QA) task, overlooking evaluation in other scenarios, such as medical language understanding and generation (Thirunavukarasu et al., 2023; He et al., 2023; Zhou et al., 2023a). This limited evaluation hinders a thorough understanding of LLM ability in diverse healthcare applications.

• (ii) Limited metric: Existing works primarily utilize matching-based metrics (e.g., Accuracy and F1) to evaluate LLM performance. These metrics fail to assess important attributes in generated responses, such as reliability and trustworthiness, which are of paramount importance for clinicians and in regulatory approvals that are essential for reliable deployment in clinical practice (Shen et al., 2023; Kitamura, 2023).

• (iii) Limited comparison: Existing works mainly compare LLM performance with their own basic models or use private datasets for evaluation (Tian et al., 2024). Such an ap-
proach falls short of providing a thorough comparative analysis among different LLMs under standardized conditions. Consequently, it hampers a comprehensive understanding of the distinct advantages and limitations of various LLMs in healthcare.

As a result, the accuracy, generalizability, and reliability of existing LLMs in diverse healthcare applications remain unclear. In response, (i) we construct the BenchHealth from the representative public health data to benchmark LLMs in healthcare. As shown in Table 1, BenchHealth encompasses three different evaluation scenarios (i.e., reasoning, generation, and understanding) and includes seven popular downstream tasks and thirteen representative datasets; Previous popular benchmarks, e.g., BLUE (Peng et al., 2019) and BLURB (Gu et al., 2021), only include the medical language understanding and close-ended question answering. (ii) In addition to the commonly used matching-related metrics, as shown in Table 2, we design additional metrics to provide insights into the reliability of LLMs in clinical settings, i.e., analyzing their ability to provide faithfulness, comprehensive, generalized, and robust information; (iii) As shown in Table 3, we collect sixteen representative LLMs that vary in the number of model parameters and structural designs. We evaluate their performance on BenchHealth for a comprehensive comparison.

The main insights from our experiments are:

- **Commercial LLMs vs. Public LLMs:** Closed-source commercial LLMs, especially GPT-4, outperform all existing open-source public LLMs on all tasks and datasets.

- **LLMs vs. State-of-the-art:** All LLMs have a strong reasoning ability to predict accurate answers from the provided options, but perform very poorly in open-ended questions, language generation, and language understanding tasks (i.e., there are significant gaps between the state-of-the-art and LLM performance).

- **Medical LLMs vs. General LLMs:** Fine-tuning general LLMs on medical data to obtain medical LLMs can improve the reasoning and understanding of medical data, but could decrease the summarization ability of LLMs.

- **Model parameters:** A larger number of model parameters can clearly improve performance on all tasks, datasets, and metrics.

- **Few-shot learning:** It leads to significant improvements in performance on medical language reasoning and generation tasks, but impairs performance on understanding tasks. On reasoning tasks, 1-shot or 3-shot learning performs the best; more examples do not lead to further improvements. On generation tasks, more samples lead to better performance.

- **Clinical usefulness:** Medical LLMs can provide more faithful answers than general LLMs (avoiding misdiagnosis) and generalize well to diverse medical tasks; General LLMs can provide more comprehensive answers than medical LLMs, which may be due to “hallucinations”, thus avoiding missed diagnoses; General LLMs have better robustness and can therefore better understand a variety of diverse inputs compared to medical LLMs.

Overall, our results show that among all types of tasks, the close-ended QA task is the only type of task in which current LLMs are comparable to state-of-the-art models and human experts. However, real-world open clinical practice diverges far from the structured nature of exam-taking. Clinical decisions, such as diagnosis and treatment recommendations, are often confronted with open-ended questions that lack pre-determined answer choices. This paradigm shift from a controlled test environment to the unpredictable and subtle domain of patient care challenges the conventional approach, demanding a more sophisticated understanding and application of medical knowledge. Our results also demonstrate that all LLMs display insufficient performance on crucial metrics necessary for ensuring the trustworthiness of LLMs in clinical settings. This unsatisfactory performance suggests that the current state of LLMs falls short of readiness for deployment in clinical settings to aid healthcare professionals. We hope that this work can offer a holistic view of LLMs in healthcare, aiming to bridge the current gaps and advance the integration of LLMs in clinical applications.

2 Benchmark

Our benchmark is shown in Table 1.

2.1 Medical Language Reasoning

We include the question answering and treatment recommendation tasks in our benchmark.
Question Answering aims to predict the correct answer to the given question. For example, the model should answer ‘D’ to the question: “Which of the following conditions does not show multifactorial inheritance? (A) Pyloric stenosis (B) Schizophrenia (C) Spina bifida (neural tube defects) (D) Marfan syndrome”. Thus, QA evaluates the correctness of the medical knowledge learned by LLMs. We include four popular datasets, i.e., MedQA (USMLE) (Jin et al., 2021), MedMCQA (Pal et al., 2022), MMLU-Medicine (Hendrycks et al., 2020), PubMedQA (Jin et al., 2019).

Treatment Recommendation is an open-ended complex task and requires the models to first understand the real-world patient-clinician conversations, in which the conversation describes the conditions and symptoms, and then recommend all possible drugs for the treatment of patients. We use ChatDoctor (Li et al., 2023b) for evaluation.

2.2 Medical Language Generation

We evaluate two popular generation tasks, i.e., radiology report summarization and discharge instruction generation.

Radiology Report Summarization aims to distill a concise summary ‘Impression’ from the lengthy ‘Findings’ section in a radiology report. ‘Findings’ contains detailed abnormal and normal clinical findings from radiology images like X-rays, CT scans, or MRI scans, and ‘Impression’ highlights the key diagnostic information and significant results, which are critical for accurate diagnosis and treatment (Jing et al., 2018; Liu et al., 2021b). We adopt the widely-used datasets, MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016).

Discharge Instruction Generation aims to generate a discharge instruction according to the patient’s health records during hospitalization when a patient is discharged from the hospital. The discharge instruction should consider diagnosis, medication, and procedure, e.g., demographics, laboratory results, admission notes, nursing notes, radiology notes, and physician notes (Liu et al., 2022). It contains multiple instructions to help the patient or carer to manage their conditions at home. We follow previous works (Liu et al., 2022) to use the MIMIC-III (Johnson et al., 2016) for evaluation.

2.3 Medical Language Understanding

We include three representative tasks, i.e., named entity extraction, relation extraction, and document classification, into our benchmark.

Named Entity Extraction can help organize and manage patient data (Perera et al., 2020). For example, it can extract medical entities mentioned in clinical notes and classify them according to relevant symptoms, medication, dosage, and procedures (Song et al., 2021). We adopt two representative datasets BC5-disease (Li et al., 2016) and NCBI-Disease (Doğan et al., 2014) for evaluation.

Relation Extraction requires the model to identify the relation between medical entities. The extracted relations provide a solid basis to link the entities in a structured knowledge base or a standardized terminology system, e.g., SNOMED CT (Chang and Mostafa, 2021; Donnelly et al., 2006) and UMLS (Bodenreider, 2004), which is critical in clinical decision support systems. We employ

Table 1: Overview of the benchmark BenchHealth for evaluating LLMs in healthcare.
As shown in Table 2, we use five metrics to benchmark LLMs in healthcare.

**Matching** We follow the common practice to calculate the classification accuracy, F1 score, ROUGE-L (Lin, 2004), and BLEU-4 (Papineni et al., 2002) to report the matching performance. Details of used metrics for different tasks are shown in Table 1. However, matching-based metrics are not specialized for evaluating the usefulness of the LLMs in clinical practice. To assist clinicians, it is necessary to provide faithful, comprehensive, and robust content (Thirunavukarasu et al., 2023; Arora and Arora, 2023; Safranek et al., 2023).

**Faithfulness** LLMs are susceptible to “hallucinations” (Li et al., 2023a; Ji et al., 2023), i.e., fluent content that appears credible but factually incorrect or potentially harmful. Therefore, it is crucial to ensure that LLMs generate faithful content, so that the models do not generate contents that “do not exist” according to clinicians (Liu et al., 2022). For instance, if clinicians annotate the ground truth contents as [Content_A, Content_B], but the model generates [Content_A, Content_C], it becomes evident that the model has introduced ‘Content_C’, which does not exist in the annotations. Such inaccuracies could lead to misdiagnoses, particularly with clinicians who have less experience. We notice that the precision scores can measure the rates of such generated non-existent content. To this end, we calculate and sum the precision scores of tasks to measure the ‘faithfulness’ scores.

**Comprehensiveness** Given the ground truth contents [Content_A, Content_B], generating comprehensive content [Content_A, Content_B] diminishes the chance of leaving out important content. They can also be used to alert clinicians to avoid missed diagnoses, improving precision medicine. The recall score measures the percentage of generated accurate content out of all correct answers. Therefore, to evaluate the comprehensiveness of model-generated contents, we calculate and sum the recall scores of different tasks to measure the ‘comprehensiveness’ scores.

**Robustness** Clinicians may express the same texts, questions, and conditions using varying formats and terminologies. For example, in the radiology report summarization task, both ‘enlargement of the cardiac silhouette’ and ‘the heart size is enlarged’ express the condition ‘cardiomegaly’. Therefore, the model needs to accurately identify ‘cardiomegaly’ for both these two different inputs. As shown in Table 1, for the report summarization task, we can compute the variance in model performance on the two datasets, IU-Xray and MIMIC-CXR (collected from different hospitals and regions, thus having different expression habits), to obtain the robustness of the model on this task. As a result, to measure the robustness scores of the

---

**Table 2:** Metrics used in our work for evaluation.

<table>
<thead>
<tr>
<th>Metrics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Matching</strong></td>
<td>(Accuracy, F1, ROUGE-L, BLEU-4)</td>
<td>Measure the match between the generated content and the ground truth content.</td>
</tr>
<tr>
<td><strong>Faithfulness</strong></td>
<td>The model cannot generate content that appears reasonable but is factually incorrect and sometimes even harmful, thus avoiding misdiagnosis.</td>
<td></td>
</tr>
<tr>
<td><strong>Comprehensiveness</strong></td>
<td>The model cannot leave out the important content, which can be used to alert clinicians to avoid missed diagnoses.</td>
<td></td>
</tr>
<tr>
<td><strong>Robustness</strong></td>
<td>For the same scenario and task, the model provides consistent and reliable performance across different formats/types/terminologies of input data (instead of overfitting specific data), measuring model stability to a range of inputs.</td>
<td></td>
</tr>
<tr>
<td><strong>Generalizability</strong></td>
<td>The model should maintain competitive performance across different scenarios and tasks (not limited to QA), to effectively assist clinicians.</td>
<td></td>
</tr>
</tbody>
</table>

---

**Table 3:** We collect 16 LLMs, including 9 general LLMs and 7 medical LLMs, covering both open-source public LLMs and closed-source commercial LLMs (gray-colored), across different numbers of parameters from 6 billion to 70 billion, and different model backbones (GLM and GPT).

<table>
<thead>
<tr>
<th>Types</th>
<th>Methods</th>
<th># Params</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General LLMs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatGLM (Tinghuai KEG, 2023)</td>
<td>6B</td>
<td></td>
</tr>
<tr>
<td>Alpaca (Taoi et al., 2023)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>Vicuna (Chiang et al., 2023)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-7B (Touvron et al., 2023c)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-13B (Touvron et al., 2023c)</td>
<td>13B</td>
<td></td>
</tr>
<tr>
<td>LLaMA-2-70B (Touvron et al., 2023c)</td>
<td>70B</td>
<td></td>
</tr>
<tr>
<td><strong>Medical LLMs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChatGLM-Med (Wang et al., 2023b)</td>
<td>6B</td>
<td></td>
</tr>
<tr>
<td>DoctorGLM (Xiong et al., 2023)</td>
<td>6B</td>
<td></td>
</tr>
<tr>
<td>Huatuor (Zhang et al., 2023a)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>ChatDoctor (Li et al., 2023b)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>Baize-Healthcare (Xu et al., 2023)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>MedAlpaca-7B (Han et al., 2023)</td>
<td>7B</td>
<td></td>
</tr>
<tr>
<td>MedAlpaca-13B (Han et al., 2023)</td>
<td>13B</td>
<td></td>
</tr>
</tbody>
</table>
Table 4: The prompts used for different evaluation tasks and datasets. We collect optimal prompts from existing state-of-the-art work.

<table>
<thead>
<tr>
<th>Prompts</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedQA (USSMLE), MedMQA, MMLU-Medicine</td>
<td>(Singhal et al., 2023b)</td>
</tr>
<tr>
<td>MedQA</td>
<td>(Singhal et al., 2023b)</td>
</tr>
<tr>
<td>PubMedQA</td>
<td>(Singhal et al., 2023b)</td>
</tr>
<tr>
<td>ChatDoctor</td>
<td>(Zhao et al., 2023b)</td>
</tr>
<tr>
<td>MIMIC-CXR, IU-Xray</td>
<td>(Tu et al., 2023)</td>
</tr>
<tr>
<td>MIMIC-III</td>
<td>(Tu et al., 2023)</td>
</tr>
<tr>
<td>BC5-disease, NCBI-Disease</td>
<td>(Chua et al., 2023)</td>
</tr>
<tr>
<td>GDD</td>
<td>(Chua et al., 2023)</td>
</tr>
<tr>
<td>GAD</td>
<td>(Chua et al., 2023)</td>
</tr>
<tr>
<td>HuCQA</td>
<td>(Chua et al., 2023)</td>
</tr>
</tbody>
</table>

As shown in Table 3, to provide a comprehensive evaluation of LLMs in healthcare, we evaluate both the general and medical LLMs. Please refer to Zhao et al. (2023); Yang et al. (2023a) and Zhou et al. (2023a); He et al. (2023) for a detailed introduction to general LLMs and medical LLMs, respectively.

General Large Language Models

We include nine general LLMs, including three leading closed-source commercial LLMs, i.e., Claude-2 (Anthropic, 2023), GPT-3.5-turbo (OpenAI, 2023a), and GPT-4 (OpenAI, 2023c), and six open-source public LLMs, i.e., ChatGLM (Tsinghua KEG, 2023; Du et al., 2022; Zeng et al., 2022), Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and LLaMA-2-7B/13B/70B (Touvron et al., 2023c).

These general LLMs are trained on a large general-purpose corpus with more than 1T tokens (Zhao et al., 2023; Yang et al., 2023a; Zhou et al., 2023a).

Medical Large Language Models

We choose seven medical LLMs with different numbers of parameters and different types of fine-tuning data. In detail, as shown in Table 3, ChatGLM-Med (Wang et al., 2023b) and DoctorGLM (Xiong et al., 2023) are fine-tuned on the ChatGLM-6B (Tsinghua KEG, 2023; Du et al., 2022; Zeng et al., 2022) using QA pairs and dialogues, respectively. Huato (Zhang et al., 2023a), ChatDoctor (Li et al., 2023b), BaiZe-Healthcare (Xu et al., 2023), and MedAlpaca-7B/13B (Han et al., 2023) are built upon the LLaMA-series models. During fine-tuning, both Huato and MedAlpaca employ the QA pairs collected from the medical knowledge graphs and medical texts, respectively. ChatDoctor and BaiZe-Healthcare are fine-tuned on medical dialogues generated by commercial LLMs (e.g., ChatGPT).

Prompts

Prompt designs are crucial for the performance of LLMs. Therefore, to ensure LLMs achieve optimal performance across different tasks,
we use tailored prompts for each task, so that LLMs can effectively understand the task and questions. In implementation, we adopt prompts used in the current state-of-the-art work for each task in the benchmark to evaluate the LLMs. Table 4 shows the prompts we used and their references.

5 Results

5.1 Medical Language Reasoning

From Table 5, we observe that on all datasets, the three leading commercial LLMs, i.e., Claude-2, GPT-3.5-turbo, and GPT-4, significantly outperform other LLMs, general or medical. In particular, on the close-ended QA task with provided options, GPT-4 even achieves a competitive accuracy of 81.2 compared to human experts (87.0) (Wu et al., 2023). In terms of open-source public LLMs, medical LLMs, e.g., ChatGLM-Med and DoctorGLM, achieve better results than general LLMs, e.g., ChatGLM, on all datasets. It indicates that fine-tuning the general LLMs on medical data can improve their performances on reasoning tasks.

Discussion The results show that, on all close-ended QA datasets, all LLMs significantly outperform existing task-specific SOTA models, e.g., PubMedBERT (Gu et al., 2022). It proves that existing LLMs have a strong reasoning ability to give accurate answers from the options. However, on the open-ended treatment recommendation task, compared with SOTA models, all LLMs achieve poor F1 scores (<15%) on the ChatDoctor dataset. This indicates a considerable need for advancement before LLMs can be integrated into actual clinical decision-making processes.

5.2 Medical Language Generation

This application is particularly useful in reducing the heavy workload of clinicians in medical text writing. Table 5 show that, among all LLMs, GPT-4 (OpenAI, 2023c) consistently achieves the best results on all generation tasks, showcasing its exceptional capability in capturing and summarizing important clinical findings compared to other LLMs. Nonetheless, the task-specific SOTA model (Hu et al., 2022) achieves 46.1 and 67.9 ROUGE-L scores on MIMIC-CXR and IU-Xray, respectively, significantly higher than all LLMs.

Discussion On the MIMIC-CXR and IU-Xray radiology report summarization datasets, most medical LLMs that have been fine-tuned on medical data, perform worse than general LLMs. In contrast, on the discharge instruction generation task, which requires the model to understand various types of medical data to provide accurate discharge instructions, medical LLMs perform better than the general LLMs. These observations may imply that the instruction fine-tuning on medical data could decrease the summarization ability of LLMs, but improve the understanding of medical data.

5.3 Medical Language Understanding

All LLMs exhibit poor performances in this scenario, including named entity extraction, relation
Figure 1: Performance (measured by traditional matching scores) of LLMs under few (1,3,5)-shot learning setting.

extraction, and document classification tasks. For example, the best results of LLMs are achieved by GPT-4 on the BC5-Disease and NCBI-Disease datasets, with 65.7 and 55.3 F1 scores, which are significantly far from current state-of-the-art performances, i.e., 90.0 F1 score achieved by SciencBERT (Beltagy et al., 2019) and 89.4 F1 score achieved by BioBERT (Lee et al., 2020), respectively. The medical LLMs have better language understanding than general LLMs in healthcare. With the same parameters, all medical LLMs outperform the general LLMs over datasets.

Discussion The inadequate performance of all LLMs may be attributed to the missing of task-specific supervised training and thus a lack of necessary medical knowledge, such as the medical terminologies for named entity extraction, the medical relations between drugs, conditions, and symptoms for relation extraction, and the background of diseases for document classification (Chen et al., 2023). As a result, existing LLMs fail to comprehend texts that typically require extensive expert knowledge to interpret. This observation underscores the effectiveness of efficiently using clinical-standard knowledge of diseases, symptoms, and medications, to fine-tune the LLMs.

5.4 Few-shot Learning Setting

We further evaluate the performance of LLMs on the few-shot learning settings, i.e., 1-shot, 3-shot, and 5-shot learning settings. We analyze the three scenarios, i.e., reasoning, generation, and understanding. For reasoning and understanding scenarios, we calculate the average performance of all datasets under that scenario to report the performance of LLMs. For the generation scenario, since the text length of the input for the discharge report generation task is long, we do not report the few-shot learning performance on the MIMIC-III dataset. Therefore, we compute the average of the performance of the other generation datasets to obtain the generation results of the LLMs. The results are reported in Figure 1.

(a) We observe that the few-shot learning can significantly boost the performances of LLMs in language reasoning. It proves the effectiveness of few-shot learning, in which the provided examples could provide efficient knowledge of medical reasoning to reason about the correct answers. However, most LLMs achieve the best results under the 1-shot and 3-shot settings. More examples (e.g., 5 shots) may not only make it difficult for LLMs to deal with long inputs but also potentially introduce
noise into the LLMs, i.e., the provided examples may not be relevant to the input problem, thus affecting performance. As a result, providing more examples does not lead to further improvements.

(b) In text generation, few-shot learning can directly demonstrate how to capture and summarize important clinical information and provide a desirable writing style. As a result, few-shot learning can consistently and substantially improve the performance of the LLMs, with more samples leading to better performance. It proves the effectiveness of using few-shot learning to significantly boost the performance of medical text generation.

(c) However, in the case of language understanding, it clearly shows that few-shot learning impairs performance. This may be because, in language understanding tasks, the characteristics of different input data are usually very different from each other, resulting in the entities or knowledge involved in the examples often being irrelevant to the test data, making the model unable to effectively utilize the examples to improve performance.

5.5 Clinical Usefulness

In Figure 2, we report the performances of LLMs in terms of clinical usefulness.

(a) In terms of faithfulness, all medical LLMs outperform general LLMs, resulting in providing more faithful answers than general LLMs, avoiding misdiagnosis.

(b) In contrast, general LLMs demonstrate better results than medical LLMs in terms of comprehensiveness, likely due to their susceptibility to “hallucinations”, meaning the LLMs tend to generate massive content including both correct and incorrect information.

(c) In terms of generalizability, we notice that medical LLMs achieve optimal results, showing that fine-tuning using the medical data can boost the overall performance of LLMs in healthcare.

(d) The general LLMs have better robustness and achieve lower robustness values than medical LLMs. For example, ChatGLM achieves 21.1 points, lower than ChatGLM-Med (22.4) and DoctorGLM (21.3).

Discussion We hypothesize that the better comprehensiveness of the general LLMs could potentially be due to that a certain degree of hallucination may offer benefits. This hypothetical advantage might assist clinicians by providing a broader spectrum of diagnostic suggestions, which could be advantageous in the diagnosis and treatment of rare diseases. However, any content generated by LLMs must be supported by factual knowledge and evidence to provide reliable, rather than misleading, results. General LLMs have better robustness, and thus can better understand a variety of diverse inputs. We speculate that the reason may be the limited diversity of fine-tuning data and tasks used for general LLMs.
Table 6: Performance of human evaluation on our BenchHealth benchmark. We compare open-source public LLMs with three leading commercial LLMs. All values are Win+Tie rates for public LLM. Higher is better in all columns.

5.6 Human Evaluation

We invite two junior annotators (medical students) and a senior annotator (medical professor) to conduct the human evaluation. All three annotators have sufficient medical knowledge. In implementations, we follow previous works (Li et al., 2023b; Zhang et al., 2023b) to randomly select 200 real patient-doctor conversations from Li et al. (2023b). We require the LLMs to simulate a doctor and provide responses based on various patient inquiries. Each junior annotator is assigned to independently compare the responses from public LLMs and those from the leading commercial LLMs, i.e., Claude-2, GPT-3.5-turbo, and GPT-4, in terms of the perceived quality of the responses. It includes faithfulness, comprehensiveness, generalizability, and robustness. The senior annotator re-evaluates the cases that are difficult for junior annotators to decide. The annotators are unaware of which model generates these reports. We report the results (win+tie rates) of public LLMs in Table 6.

We observe that with the same number of model parameters, medical LLMs outperform general LLMs in terms of faithfulness and generalizability, but underperform general LLMs in comprehensiveness and robustness. These results are consistent with those shown in Figure 2, which demonstrates the validity and appropriateness of our metric and benchmark.

6 Conclusions

This paper introduces BenchHealth, a healthcare benchmark encompassing medical language reasoning, generation, and comprehension scenarios. It employs metrics that extend beyond mere accuracy, aiming to evaluate the utility and reliability of LLMs for clinical applications. Although LLMs have made promising advances, our analysis uncovers a gap between the capabilities of LLMs and the requirements for clinical application, especially in open-ended non-QA tasks that lack pre-determined answer choices, underscoring the challenges LLMs face in providing reliable support in healthcare.

Limitations

A limitation of this work is that the recent development of LLMs is rapid and we do not evaluate the latest LLMs, e.g., GPT-4.5 and Qwen (Bai et al., 2023), and medical LLMs, e.g., Zhongjing (Yang et al., 2023b) and Qilin-Med (Ye et al., 2023).

References


Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.


Conrad W Safranek, Anne Elizabeth Sidamon-Eristoff, Aidan Gilson, and David Chartash. 2023. The role of large language models in medical education: applications and implications.


Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023a. Harnessing the power of llms in


