Large Language Models in Healthcare: A Comprehensive Benchmark

Fenglin Liu[1], Hongjian Zhou[1], Yining Hua[2], Omid Rohanian[1], Lei Clifton[3], David A. Clifton[1]
[1]Institute 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 (Singhal et al., 2023a,b), resulting in ChatDoctor (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., 2023b) 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.

Scenarios	Tasks	Datasets	Data Domains	Sizes	Matching Metrics
Medical Language Reasoning	Question Answering	MedQA (USMLE) (Jin et al., 2021)	Medical Licensing Examination	1,273	Accuracy
		MedMCQA (Pal et al., 2022)	Medical Entrance Examination	4,183	Accuracy
		MMLU-Medicine (Hendrycks et al., 2020)	Professional&College Medicine	272	Accuracy
		PubMedQA (Jin et al., 2019)	Medical Literature	500	Accuracy
	Treatment Recommendation	ChatDoctor (Li et al., 2023b)	Patient-Clinician Conversations	796	Micro F1
Medical Language Generation	Radiology Report Summarization	MIMIC-CXR (Johnson et al., 2019)	Radiography	3,269	ROUGE-L
		IU-Xray (Demner-Fushman et al., 2016)	Radiography	341	ROUGE-L
	Discharge Instruction Generation	MIMIC-III (Johnson et al., 2016)	Critical Care	3,633	BLEU-4
Medical Language Understanding	Named Entity Recognition	BC5-disease (Li et al., 2016)	Scientific Literature	4,797	F1 entity-level
		NCBI-Disease (Doğan et al., 2014)	Scientific Literature	940	F1 entity-level
	Relation Extraction	DDI (Segura-Bedmar et al., 2013)	Drug	5,716	Micro F1
		GAD (Becker et al., 2004)	Genetic	534	Micro F1
	Document Classification	HoC (Baker et al., 2016)	Scientific Literature	315	Micro F1

Table 1: Overview of the benchmark *BenchHealth* for evaluating LLMs in healthcare.

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

Metrics
Matching (Accuracy, F1, ROUGE-L, BLEU-4) Measure the match between the generated content and the ground truth content.
Faithfulness The model can not generate content that appears reasonable but is factually incorrect and sometimes even harmful, thus avoiding misdiagnosis.
Comprehensiveness The model can not leave out the important content, which can be used to alert clinicians to avoid missed diagnoses.
Robustness 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.
Generalizability The model should maintain competitive performance across different scenarios and tasks (not limited to QA), to effectively assist clinicians.

Table 2: Metrics used in our work for evaluation.

Types	Methods	# Params
General LLMs	Claude-2 (Anthropic, 2023)	-
	GPT-3.5-turbo (OpenAI, 2023a)	-
	GPT-4 (OpenAI, 2023c)	-
	ChatGLM (Tsinghua KEG, 2023)	6B
	Alpaca (Taori et al., 2023)	7B
	Vicuna (Chiang et al., 2023)	7B
	LLaMA-2-7B (Touvron et al., 2023c)	7B
	LLaMA-2-13B (Touvron et al., 2023c)	13B
	LLaMA-2-70B (Touvron et al., 2023c)	70B
Medical LLMs	ChatGLM-Med (Wang et al., 2023b)	6B
	DoctorGLM (Xiong et al., 2023)	6B
	Huatuo (Zhang et al., 2023a)	7B
	ChatDoctor (Li et al., 2023b)	7B
	Baize-Healthcare (Xu et al., 2023)	7B
	MedAlpaca-7B (Han et al., 2023)	7B
	MedAlpaca-13B (Han et al., 2023)	13B

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).

the DDI (Segura-Bedmar et al., 2013) and GAD (Becker et al., 2004) to evaluate LLMs.

Document Classification is a document-level language understanding task aiming to predict multiple correct labels to the input medical document, and can be used to improve clinical management systems. We use the representative dataset HoC (Baker et al., 2016) for evaluation.

3 Metrics

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

Prompts	Sources
MedQA (USMLE), MedMCQA, MMLU-Medicine The following are multiple-choice questions about medical knowledge. Solve them in a step-by-step fashion, starting by summarizing the available information. Output a single option from the four options as the final answer.	(Singhal et al., 2023b)
PubMedQA This is a multiple-choice question about medical research. Determine the answer to the question based on the strength of the scientific evidence provided in the context. Valid answers are yes, no, or maybe. Answer yes or no if the evidence in the context supports a definitive answer. Answer maybe if the evidence in the context does not support a definitive answer, such as when the context discusses both conditions where the answer is yes and conditions where the answer is no.	(Singhal et al., 2023b)
ChatDoctor "task": "Your task is to list the medications based on the provided content related to the symptom or disease mentioned in the question. Understand the question, extract relevant information, analyze it, and provide a concise and accurate answer.", "answer format": Analysis: Provide an analysis that logically leads to the answer based on the relevant content. Final Answer: Provide the final answer, which should be a list of medications related to the symptom or disease. "not to dos": "Do not make assumptions not supported by the content. Avoid providing personal opinions or interpretations. Summarize and interpret the information as objectively and accurately as possible. You are providing an analysis, not diagnosing or treating medical conditions."	(Zhou et al., 2023b)
MIMIC-CXR, IU-Xray You are a helpful radiology assistant. The following are questions about radiology reports. Summarize the findings in the report into diagnostic statements in a coherent paragraph. Given the findings: {Findings}. Q: Summarize the findings. A:	(Tu et al., 2023)
MIMIC-III Provide plain language discharge instructions, containing the following three main components from patients' perspective: (1) What is my main health condition? (i.e., why was I in the hospital?) (2) What do I need to do? (i.e., how do I manage at home, how should I best care for myself, what medications to take, and which appointments to go to next (if available)) (3) Why is it important for me to do this?	(Fleming et al., 2023)
BC5-disease, NCBI-Disease Paragraph: <Paragraph ID> | <text> Please extract all chemicals/genes/diseases mentioned in the paragraph. Answer with the format "<Paragraph ID> | <recognized entities>"	(Chen et al., 2023)
DDI @DRUG$ an anionic-binding resin, has a considerable effect in lowering the rate and extent of @DRUG$ bioavailability. Target: You need to identify the relationship between the two @DRUG$. Require: you must start with choose one from the ["mechanism," "effect," "advice," "int," "None"], Specific Explanation: mechanism: This type is used to annotate DDIs that are described by their PK mechanism (e.g. Grepafloxacin may inhibit the metabolism of theobromine). effect: This type is used to annotate DDIs describing an effect (e.g. In uninfected volunteers, 46% developed rash while receiving SUSTIVA and clarithromycin) or a PD mechanism (e.g. Chlorthali done may potentiate the action of other antihypertensive drugs). advice: This type is used when a recommendation or advice Regarding a drug interaction is given (e.g. UROXATRAL should not be used in combination with other alpha-blockers). int: This type is used when a DDI appears in the text without providing any additional information (e.g. the interaction of Omeprazole and ketoconazole have been established). You should mark the final category with < >.	(Chen et al., 2023)
GAD Given a sentence that introduces a gene (denoted as "@GENE$") and a disease (denoted as "@DISEASE$"), predict whether the gene and disease have a relation or not. The relation between the gene and disease can be any functional, causal, or associative connection. If there is a relation, then the label should be "Yes", otherwise "No".	(Tang et al., 2023)
HoC document: < text>; target: The correct category for this document is ? You must choose from the given list of answer categories (introduce what each category is ...)"	(Chen et al., 2023)

Table 4: The prompts used for different evaluation tasks and datasets. We collect optimal prompts from existing state-of-the-art work.

language reasoning, generation, and understanding scenarios, we respectively calculate the variance in model performance on the representative question answering, radiology report summarization, and named entity recognition tasks. Finally, we sum the variance up to obtain the overall robustness scores of the LLMs, reflecting whether their accuracy is significantly impacted by variations in the inputs.

Generalizability To effectively support clinicians in different settings, LLMs should perform well in a wide range of scenarios and tasks (not limited to QA). For clarity, we directly average all the matching scores to obtain the 'generalizability' scores to evaluate how LLMs perform across a range of scenarios and tasks.

4 Large Language Models

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. *Huatuo* (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 Huatuo 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,

Types	Methods	# Params	Language Reasoning					Language Generation			Language Understanding				
			MedQA	MedMCQA	MMLU	PubMedQA	ChatDoc.	MIMIC-CXR	IU-Xray	MIMIC-III	BC5	NCBI	DDI	GAD	HoC
Task-specific SOTA		-	44.6	43.0	-	60.2	-	**46.1**	**67.9**	**30.5**	**90.0**	**89.4**	**84.1**	**84.0**	**85.1**
General LLMs	Claude-2	-	65.1	60.3	78.7	70.8	9.1	13.3	9.4	26.1	52.9	44.2	50.4	50.7	70.8
	GPT-3.5-turbo	-	61.2	59.4	73.5	70.2	7.3	14.1	10.3	28.6	52.3	46.1	49.3	50.8	66.4
	GPT-4	-	**81.2**	**74.6**	**90.8**	**76.6**	**13.7**	15.2	11.4	30.1	65.7	55.3	62.6	64.4	78.1
	ChatGLM	6B	25.7	24.2	33.5	53.0	2.9	13.3	7.5	18.6	37.2	31.9	34.1	36.6	47.5
	Alpaca	7B	34.2	30.1	40.8	65.2	3.5	12.6	8.7	20.4	41.2	36.5	37.4	36.9	52.6
	Vicuna	7B	34.5	33.4	43.4	64.8	2.6	13.8	8.2	23.4	44.5	37.0	39.4	41.2	53.8
	LLaMA-2-7B	7B	32.9	30.6	42.3	63.4	3.3	12.3	8.6	20.2	40.1	34.8	37.9	39.3	48.6
	LLaMA-2-13B	13B	38.1	35.5	46.0	66.8	4.8	12.0	9.1	21.1	46.6	38.3	39.7	41.2	55.9
	LLaMA-2-70B	70B	45.8	42.7	54.0	67.4	5.5	13.9	8.0	23.2	47.8	41.5	45.6	44.7	63.2
Medical LLMs	ChatGLM-Med	6B	27.3	25.8	35.3	58.8	3.3	9.5	4.7	19.4	40.5	35.2	37.4	33.6	49.3
	DoctorGLM	6B	25.9	23.1	36.8	57.4	3.1	6.5	3.5	15.2	38.7	33.6	35.6	34.7	50.8
	Huatuo	7B	28.4	24.8	31.6	61.0	3.8	8.7	3.8	17.8	43.6	37.5	40.1	38.2	50.2
	ChatDoctor	7B	33.2	31.5	40.4	63.8	5.3	8.9	4.2	20.7	45.8	40.9	41.2	40.1	55.7
	Baize-Healthcare	7B	34.9	31.3	41.9	64.4	4.7	9.8	4.4	19.3	44.4	38.5	41.9	45.8	54.5
	MedAlpaca-7B	7B	35.1	32.9	48.5	62.4	4.8	10.4	7.6	22.1	47.3	39.0	43.5	44.0	58.7
	MedAlpaca-13B	13B	37.3	35.7	51.5	65.6	5.1	11.7	8.6	24.7	49.2	41.6	44.1	44.5	59.4

Table 5: Performance (measured by traditional matching scores) of LLMs under the zero-shot learning setting. We denote the results of three commercial LLMs (gray-colored) as upper bounds on the performance of open-source public LLMs. For comparison, in the first row, we also report the results of task-specific state-of-the-art (SOTA) models, which are fine-tuned in a fully supervised manner on downstream data and tasks. The close-ended QA task is the only task for which the current LLMs are comparable to the SOTA.

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

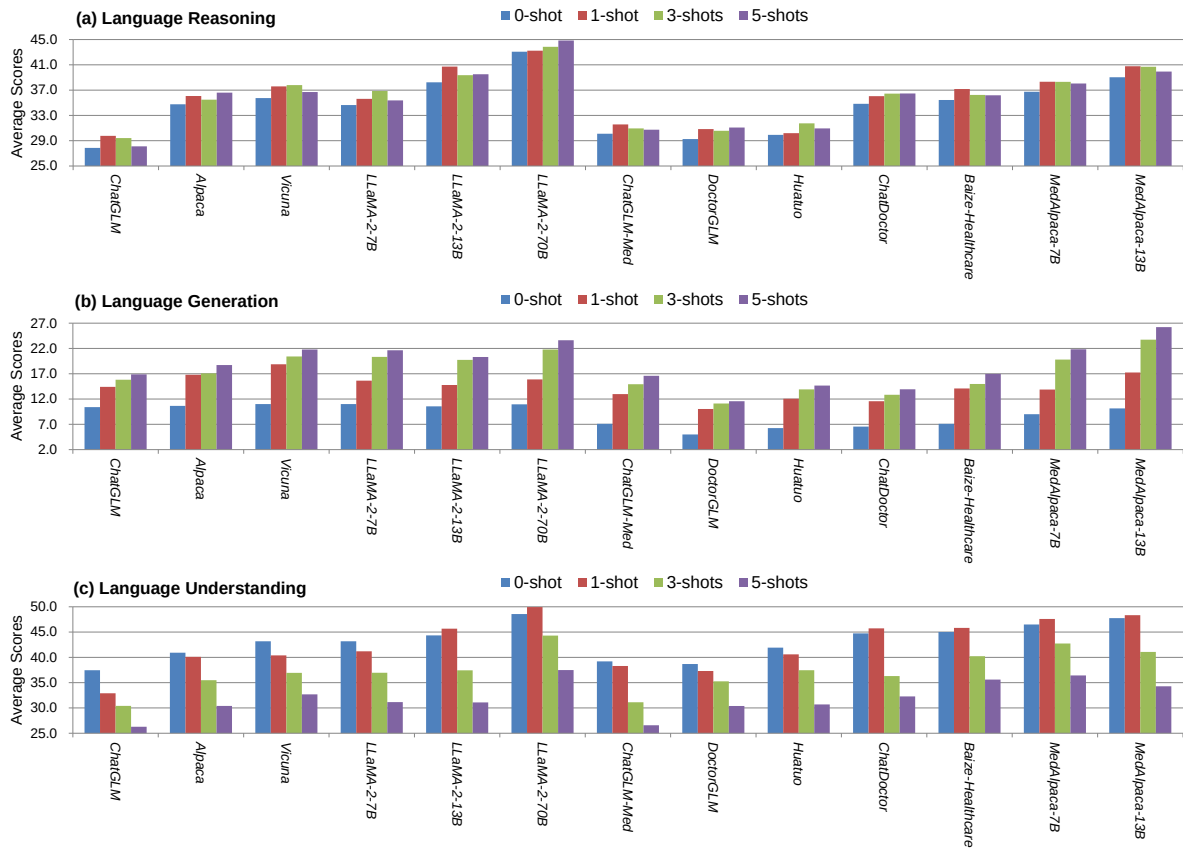

Figure 1: Performance (measured by traditional matching scores) of LLMs under few (1,3,5)-shot learning setting.

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 SciBERT (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

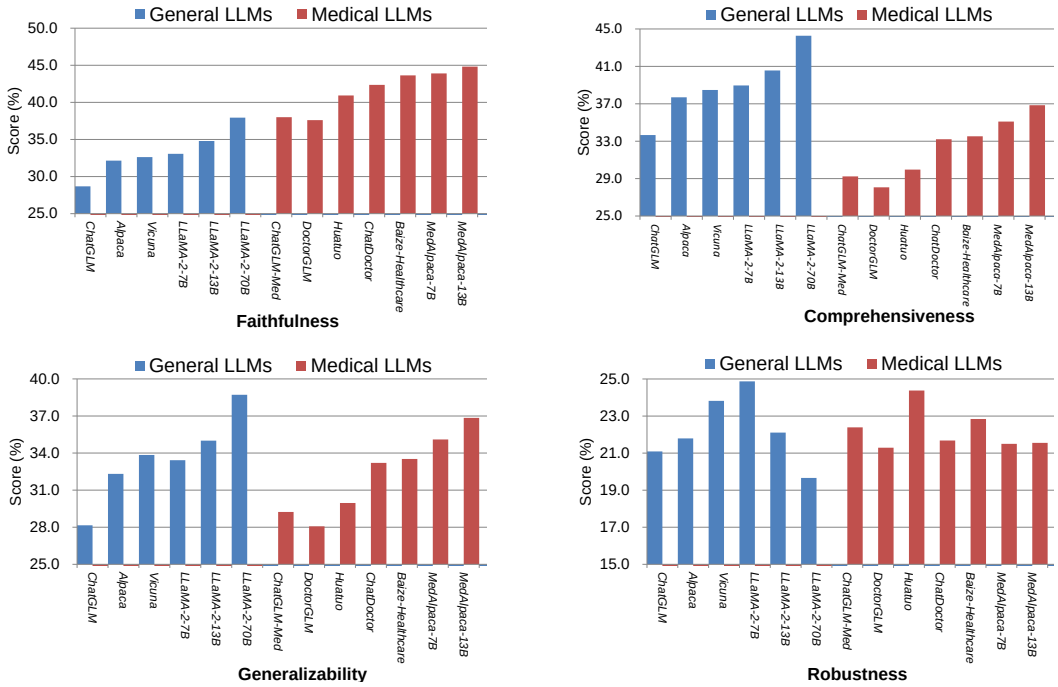

Figure 2: Performance of existing LLMs on our *BenchHealth* benchmark in terms of clinical usefulness. Higher faithfulness, comprehensiveness, and generalizability scores are better. Lower robustness scores are better.

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

Types	Methods	# Params	Faithfulness			Comprehensiveness			Generalizability			Robustness		
			Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4
General LLMs	Alpaca	7B	29.5	35.0	9.5	29.0	40.0	18.0	21.0	26.0	17.5	32.0	42.5	23.5
	Vicuna	7B	34.0	39.5	14.0	33.5	43.5	30.5	35.5	41.0	22.0	39.0	37.5	20.0
	LLaMA-2-7B	7B	32.5	37.0	13.0	40.5	48.0	26.5	29.0	34.5	21.5	44.5	46.5	27.5
	LLaMA-2-13B	13B	39.5	44.0	16.0	47.0	52.5	37.0	43.0	49.0	32.5	52.5	49.5	31.0
	LLaMA-2-70B	70B	43.0	49.5	19.5	**52.5**	**58.0**	**41.5**	**54.5**	**56.5**	**39.0**	**58.5**	**61.0**	**45.0**
Medical LLMs	ChatDoctor	7B	38.0	46.5	23.0	18.0	16.5	8.0	25.0	27.0	15.5	20.0	24.5	11.0
	Baize-Healthcare	7B	44.5	52.0	28.5	29.5	33.5	18.5	39.5	45.5	28.0	36.0	42.0	22.5
	MedAlpaca-7B	7B	47.0	55.5	31.5	26.0	33.0	15.5	33.5	31.0	19.0	30.5	37.5	17.0
	MedAlpaca-13B	13B	**50.5**	**61.0**	**34.0**	31.0	35.5	19.5	38.5	39.0	20.5	37.5	43.0	24.0

Table 6: Performance of human evaluation on our *BenchHealth* benchmark. We compare open-source public LLM with three leading commercial LLMs. All values are Win+Tie rates for public LLM. Higher is better in all columns.

to develop medical LLMs (Rohanian et al., 2023). It leads to overfitting to specific types of data and thus reduces the robustness of the model during instruction fine-tuning.

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

Anthropic. 2023. Claude-2.

Anmol Arora and Ananya Arora. 2023. The promise of large language models in health care. *The Lancet*, 401(10377):641.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. 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, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Simon Baker, Ilona Silins, Yufan Guo, Imran Ali, Johan Högberg, Ulla Stenius, and Anna Korhonen. 2016. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440.

Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. 2004. The genetic association database. *Nature genetics*, 36(5):431–432.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems*.

Eunsuk Chang and Javed Mostafa. 2021. The use of snomed ct, 2013-2020: a literature review. *Journal of the American Medical Informatics Association*, 28(9):2017–2026.

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Medical Informatics Assoc.*, 23(2):304–310.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

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.

Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo P Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. 2023. Medalign: A clinician-generated dataset for instruction following with electronic medical records. *arXiv preprint arXiv:2308.14089*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

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*.

Jinpeng Hu, Zhihong Chen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2022. Improving radiology summarization with radiograph and anatomy prompts. *arXiv preprint arXiv:2210.08303*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In *Annual Meeting of the Association for Computational Linguistics*.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Mahdi Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Felipe C Kitamura. 2023. Chatgpt is shaping the future of medical writing but still requires human judgment. *Radiology*, page 230171.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints*, pages arXiv–2305.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *ACL*.

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *Annual Meeting of the Association for Computational Linguistics*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. 2022. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. *Advances in Neural Information Processing Systems*, 35:18864–18877.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. 2021c. Auto-encoding knowledge graph for unsupervised medical report generation. In *Advances in Neural Information Processing Systems*.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2023a. Chatgpt [large language model]. https://chat.openai.com.

OpenAI. 2023b. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenAI. 2023c. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for automatic evaluation of machine translation. In *ACL*.

Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Sajan B Patel, Kyle Lam, and Michael Liebrenz. 2023. Chatgpt: friend or foe. *Lancet Digit. Health*, 5:e102.

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 *BioNLP@ACL*, pages 58–65.

Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. 2020. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673.

Omid Rohanian, Mohammadmahdi Nouriborji, and David A Clifton. 2023. Exploring the effectiveness of instruction tuning in biomedical language processing. *arXiv preprint arXiv:2401.00579*.

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.

Isabel Segura-Bedmar, Paloma Martínez Fernández, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. Chatgpt and other large language models are double-edged swords. *Radiology*, page 230163.

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. 2023a. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023b. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023c. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. 2021. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):bbab282.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023b. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023c. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tsinghua KEG. 2023. Chatglm-6b: A large-scale language model. https://github.com/THUDM/ChatGLM-6B/blob/main/README_en.md. Accessed: 2023-11-05.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. 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. *arXiv preprint arXiv:2304.06975*.

Haochun Wang, Chi Liu, Sendong Zhao, Bing Qin, and Ting Liu. 2023b. Chatglm-med. https://github.com/SCIR-HI/Med-ChatGLM.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

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*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

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

practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023b. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*.

Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, and Andrew Liu. 2023. Qilinmed: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023b. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, et al. 2023a. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Hongjian Zhou, Fenglin Liu, Wenjun Zhang, Guowei Huang, Lei Clifton, David Eyre, Haochen Luo, Fengyuan Liu, Kim Branson, Patrick Schwab, et al. 2023b. Druggpt: A knowledge-grounded collaborative large language model for evidence-based drug analysis. *Preprint*.