Large Language Models in the Clinic: A Comprehensive Benchmark

Fenglin Liu[1], Hongjian Zhou[1], Yining Hua[2], Omid Rohanian[1], Anshul Thakur[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 (QA) task with answer options for evaluation. However, many clinical decisions involve answering open-ended questions without pre-set options. To better understand LLMs in the clinic, we construct a benchmark *ClinicBench*. We first collect eleven existing datasets covering diverse clinical language generation, understanding, and reasoning tasks. Furthermore, we construct six novel datasets and complex clinical tasks that are close to real-world practice, i.e., referral QA, treatment recommendation, hospitalization (*long* document) summarization, patient education, pharmacology QA and drug interaction for emerging drugs. We conduct an extensive evaluation of twenty-two LLMs under both zero-shot and few-shot settings. Finally, we invite medical experts to evaluate the clinical usefulness of LLMs.

1 Introduction

Large language models (LLMs), such as ChatGPT (OpenAI, 2023b), are increasingly being recognized for their potential in healthcare to aid clinical decision-making. Many efforts have been made to develop medical LLMs (Singhal et al., 2023a,c; Zhou et al., 2023). Existing 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., 2023b) on the United States Medical Licensing Examination (USMLE).

Despite the promising results of existing LLMs, several issues need to be addressed for better use of LLMs in assisting clinicians: **(i) Limited evaluation:** Most works only focus on evaluating LLMs in close-ended (exam-style) QA tasks, overlooking their evaluation in other scenarios, such as clinical language understanding and generation (He et al., 2023); **(ii) Limited task:** Current works primarily focus on non-clinical machine learning tasks, which cannot adequately evaluate the models' ability to solve complex clinical problems, e.g., health education (Safranek et al., 2023), treatment recommendation (Wilhelm et al., 2023), and emerging drug analysis; **(iii) Limited comparison:** Existing works either provide limited qualitative examples (Patel and Lam, 2023) or use limited baselines (mainly focusing on ChatGPT) for quantitative comparisons (Chen et al., 2023a,b; Jahan et al., 2024). Therefore, existing works fall short of providing a thorough comparative analysis of different LLMs across diverse clinical scenarios and tasks.

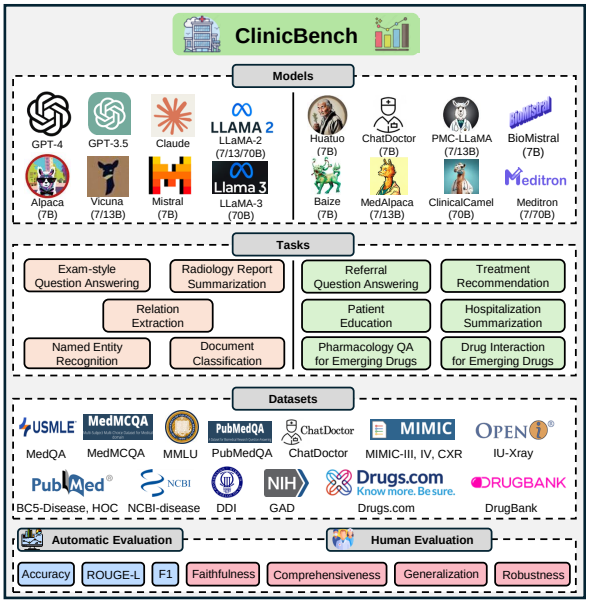

Figure 1: Overview of our ClinicBench, which includes 22 LLMs, 11 tasks, 17 datasets, and multiple metrics across automatic and human evaluations.

In this paper, as shown in Figure 1, we propose the ClinicBench, which encompasses eleven downstream tasks (across three different scenarios, i.e., reasoning, generation, and understanding) and seventeen datasets to provide a comprehensive eval-

Scenarios	Tasks	Descriptions	Datasets	Sizes	Metrics
Clinical Language Reasoning	Exam-style QA	Predict the correct answer to the given question from the provided choices.	MedQA (USMLE) (Jin et al., 2021) MedMCQA (Pal et al., 2022) MMLU-Med (Hendrycks et al., 2020) PubMedQA (Jin et al., 2019)	1,273 4,183 272 500	Accuracy Accuracy Accuracy Accuracy
	Referral QA	Predict the correct answer to the given question about patients' treatments and medications, based on their referral letters.	Derived from MIMIC-IV (Johnson et al., 2023)	1,057	Accuracy
	Treatment Recommendation	Recommend all appropriate drugs for the treatment of patients, given their conditions and symptoms.	Derived from ChatDoctor (Li et al., 2023b)	796	Micro F1
Clinical Language Generation	Radiology Report Summarization	Generate a concise 'Impression' section from the lengthy 'Findings' section in a radiology report.	MIMIC-CXR (Johnson et al., 2019) IU-Xray (Demner-Fushman et al., 2016)	3,269 341	ROUGE-L ROUGE-L
	Hospitalization Summarization	Summarize the key diagnostic information and significant results based on the patients' multiple health (*long*) records during hospitalization, e.g., physician notes, nursing notes, and medication.	Derived from MIMIC-IV (Johnson et al., 2023)	382	ROUGE-L
	Patient Education	Generate educational instructions to help patients manage their conditions, according to their health (*long*) documents.	Derived from MIMIC-III (Johnson et al., 2016)	181	ROUGE-L
Clinical Language Understanding	Named Entity Recognition (NER)	Extract medical entities mentioned in clinical notes and classify them according to relevant symptoms, medications, dosages, and procedures.	BC5-Disease (Li et al., 2016) NCBI-Disease (Doğan et al., 2014)	4,797 940	F1 entity-level F1 entity-level
	Relation Extraction (RE)	Identify the relations, e.g., the mechanism of interaction, the effect of interaction, between medical entities mentioned in the text.	DDI (Segura-Bedmar et al., 2013) GAD (Becker et al., 2004)	5,716 534	Micro F1 Micro F1
	Document Classification (DC)	Predict multiple correct labels to the input clinical document.	HoC (Baker et al., 2016)	315	Micro F1
	Pharmacology QA for Emerging Drugs	Predict the correct answer to the given pharmacology question for the new drugs released between October 2023 and April 2024.	Derived from DrugBank (Wishart et al., 2018)	213	Accuracy
	Drug Interaction for Emerging Drugs	Assess whether the therapeutic efficacy of the Moderna COVID-19 Vaccine can be decreased when used in combination with other drugs.	Derived from Drug.com (Durgs.com, 2024)	200	Accuracy

Table 1: Overview of our evaluation scenarios, which includes eleven existing datasets covering five non-clinical machine learning tasks and six novel datasets covering six complex clinical tasks (gray-highlighted text).

uation of LLMs in the clinic. Most recently, several works (Chen et al., 2023a,b) have attempted to benchmark LLMs in healthcare. However, they only adopt the non-clinical machine learning tasks from existing benchmarks (i.e., BLUE (Peng et al., 2019) and BLURB (Gu et al., 2021)) for evaluation. As shown in Table 1, we further propose six clinical tasks and build six novel datasets to evaluate the performance of LLMs in multiple complex clinical scenarios, i.e., open-ended question decision-making, long document processing, and new drug understanding, all of which are very common in the clinic. Meanwhile, it is worth noting that existing works only use three LLMs, i.e., LLaMA, GPT-3.5, and GPT-4, for evaluation. In our work, as shown in Table 2, we collect twenty-two diverse LLMs for a comprehensive comparison. In terms of evaluations, existing works mainly consider 0-shot and 1-shot experiments. We further perform 3-shot and 5-shot experiments, and provide human evaluations to offer insights into LLMs in the clinic.

Finally, to build medical LLMs, we notice that existing works (He et al., 2023) mainly adopt medical consultant dialogues, exam-style QA, and articles as fine-tuning data, which, however, cannot directly or fully represent clinical knowledge. For example, dialogues contain irrelevant and incomplete information, and articles may focus on laboratory results rather than clinical practice. Thus, we explore using clinical-standard knowledge bases collected from the clinic to develop a medical LLM. We then analyze how different types of fine-tuning data affect medical LLMs' performance. The evaluation proves the importance of introducing clinical-standard knowledge bases for fine-tuning and increasing the diversity of fine-tuning data.

The main contributions of this work are:

- We construct the *ClinicBench* with three scenarios, eleven tasks, and seventeen datasets, to benchmark twenty-two LLMs under both zero-shot and few-shot settings.
- We build six novel datasets tailored to clinical practice to measure the capabilities of LLMs in solving complex clinical problems. We further perform human evaluation to benchmark the clinical usefulness of LLMs.
- We conduct a preliminary exploration of using clinical-standard knowledge bases as the fine-tuning data to develop medical LLMs and analyze the effect of the fine-tuning data.

2 ClinicBench

Table 1 illustrates our benchmark. Here, we mainly introduce our built six clinical tasks and datasets. Please refer to our supplementary material and Appendix A for more details of our benchmark.

i) *Referral QA*: When a patient returns to their GP or is referred to another hospital, this task can

Types	Methods	# Params
General Large Language Models	Claude-2 (Anthropic, 2023)	Commercial
	GPT-3.5-turbo (OpenAI, 2023a)	Commercial
	GPT-4 (OpenAI, 2023c)	Commercial
	Alpaca (Taori et al., 2023)	7B
	Vicuna-7B (Chiang et al., 2023)	7B
	LLaMA-2-7B (Touvron et al., 2023b)	7B
	Mistral (Jiang et al., 2023)	7B
	Vicuna-13B (Chiang et al., 2023)	13B
	LLaMA-2-13B (Touvron et al., 2023b)	13B
	LLaMA-2-70B (Touvron et al., 2023b)	70B
	LLaMA-3-70B (Meta, 2024)	70B
Medical Large Language Models	Huatuo (Zhang et al., 2023a)	7B
	ChatDoctor (Li et al., 2023b)	7B
	PMC-LLaMA-7B (Wu et al., 2023a)	7B
	Baize-Healthcare (Xu et al., 2023)	7B
	MedAlpaca-7B (Han et al., 2023)	7B
	Meditron-7B (Chen et al., 2023c)	7B
	BioMistral (Labrak et al., 2024)	7B
	PMC-LLaMA-13B (Wu et al., 2023a)	13B
	MedAlpaca-13B (Han et al., 2023)	13B
	ClinicalCamel (Toma et al., 2023)	70B
	Meditron-70B (Chen et al., 2023c)	70B

Table 2: We collect 22 LLMs (i.e., 11 general LLMs and 11 medical LLMs) covering open-source public LLMs and closed-source commercial LLMs, across different numbers of parameters from 7 to 70 billion (B).

help clinicians quickly understand the patient's treatment and medication. We randomly extract 1,000 referral letters from MIMIC-IV (Johnson et al., 2023) and use GPT-4 to generate multiple QA pairs about medications and treatments for each letter. After review, correction, and filtering by experts, we finally obtain 1,057 QA pairs.

ii) *Treatment Recommendation*: It requires providing all possible appropriate medications for treating the current patient's condition, thus helping the clinician designate a treatment plan. We collect 796 patient-physician conversations, which describe the patient's diseases and symptoms, from ChatDoctor (Li et al., 2023b). We further collect the corresponding drug recommendations given by the physicians. During evaluation, we ask the LLMs to list all available drugs for the treatment of patients based on their diseases and symptoms.

iii) *Hospitalization Summarization*: Clinicians need to spend about 40%-50% of their time in their daily work (Sinsky et al., 2016) reading lots of patients' health documents and writing a hospitalization summary that highlights key diagnostic information, which is crucial for the patient's discharge or transfer to another hospital. To this end, we first collect 1,000 EHRs from MIMIC-IV resource (Johnson et al., 2023) and then exclude those without a discharge summary. The remaining EHRs are further reviewed and filtered by experts, resulting in 382 EHRs, which are used to construct 382 pairs of clinical documents and summaries. The average length of the input documents is 1,787 words.

iv) *Patient Education*: Similarly, clinicians need to read lots of health documents to write educational materials to guide patients on how to better manage their conditions. Therefore, a desirable patient education generation system can substantially reduce clinical workload. We adopt the MIMIC-III (Johnson et al., 2016) to collect 181 pairs of health documents and educational instructions. The average length of the input documents is 3,037 words.

v) *Pharmacology QA for Emerging Drugs:* It aims to answer pharmacology-related questions based on the given new drugs. We collect 213 new drugs released on the pharmaceutical knowledge database DrugBank (Wishart et al., 2018) between October 2023 and April 2024. Then, we use GPT-4 (OpenAI, 2023c) to generate 213 question-answer pairs, which are further reviewed by experts.

vi) *Drug Interaction for Emerging Drugs*: It aims to predict the effects of known drug combinations given the descriptions of new drugs. To this end, we randomly chose 100 identified drug interactions with the Moderna COVID-19 Vaccine (2023-2024 Formula) and 100 drugs without interaction from Drug.com (Durgs.com, 2024). The LLMs are tasked with assessing whether the therapeutic efficacy of the vaccine can be decreased when used in combination with other drugs.

Discussion Unlike exam-style QA tasks with answer options, the treatment recommendation is an open-ended task where the LLMs need to rely on their own knowledge to reason and make decisions; Both patient education and hospitalization summarization tasks require LLMs to process patient documents with lengths around 2k-3k words, and thus can evaluate LLMs' ability to deal with *long* health documents, which are common in the clinic; The tasks of pharmacology QA and drug interaction for emerging drugs are crucial for supporting the decision-making and management of new drugs, which frequently emerge in clinical practice.

3 Results

In this section, we will show the detailed results and analyses of LLMs in our benchmark.

3.1 Settings

As shown in Table 2, we collect 22 diverse LLMs to provide a comprehensive benchmark. Please refer to Zhao et al. (2023); He et al. (2023) or our Appendix for a detailed introduction to them. To

Types	Methods	# Params	Clinical Language Reasoning						Clinical Language Generation				Clinical Language Understanding						
			Exam-style QA				Referral QA	Treat Recom.	Report Summari.		Hospitaliz. Summari.	Patient Education	NER		RE		DC	Pharma. QA	Drug Inter.
			MedQA	MedMCQA	MMLU	PubMedQA			MIMIC	IU-Xray			BC5	NCBI	DDI	GAD	HoC		
Task-specific SOTA		-	44.6	43.0	-	60.2	-	-	**46.1**	**67.9**	-	-	**90.0**	**89.4**	**84.1**	**84.0**	**85.1**	-	-
General Large Language Models	Claude-2	Com.	65.1	60.3	78.7	70.8	80.5	9.1	13.3	9.4	11.3	8.4	52.9	44.2	50.4	50.7	70.8	60.6	51.5
	GPT-3.5-turbo	Com.	61.2	59.4	73.5	70.2	81.1	7.3	14.1	10.3	10.5	9.2	52.3	46.1	49.3	50.8	66.4	57.3	47.0
	GPT-4	Com.	**83.4**	**78.2**	**92.3**	**80.0**	**83.2**	**18.6**	20.7	18.6	**14.2**	**12.7**	71.3	58.4	64.6	68.2	83.6	**63.8**	**56.5**
	Alpaca	7B	34.2	30.1	40.8	65.2	74.8	3.5	12.6	8.7	4.1	2.9	41.2	36.5	37.4	36.9	52.6	41.3	47.5
	Vicuna-7B	7B	34.5	33.4	43.4	64.8	76.4	2.6	13.8	8.2	4.5	3.1	44.5	37.0	39.4	41.2	53.8	42.3	45.5
	LLaMA-2-7B	7B	32.9	30.6	42.3	63.4	74.5	3.3	12.3	8.6	4.9	4.6	40.1	34.8	37.9	39.3	48.6	46.5	48.0
	Mistral	7B	35.7	37.8	46.3	69.4	77.7	5.0	13.2	7.9	6.1	5.3	46.8	39.9	43.5	44.3	59.6	51.2	53.0
	Vicuna-13B	13B	38.0	36.4	45.6	66.2	76.8	4.6	14.5	9.4	6.2	4.7	46.2	39.0	41.3	43.5	56.7	45.1	46.0
	LLaMA-2-13B	13B	38.1	35.5	46.0	66.8	77.1	4.8	12.0	9.1	6.4	5.6	46.6	38.3	39.7	41.2	55.9	46.9	47.5
	LLaMA-2-70B	70B	45.8	42.7	54.0	67.4	78.9	5.5	13.9	8.0	8.3	6.8	47.8	41.5	45.6	44.7	63.2	49.3	51.5
	LLaMA-3-70B	70B	78.8	74.7	86.4	77.4	82.4	10.2	18.4	15.5	10.9	10.1	63.7	50.2	59.7	63.1	79.0	62.4	53.0
Medical Large Language Models	Huatuo	7B	28.4	24.8	31.6	61.0	69.3	3.8	8.7	3.8	2.2	1.4	43.6	37.5	40.1	38.2	50.2	44.1	49.5
	ChatDoctor	7B	33.2	31.5	40.4	63.8	73.7	5.3	8.9	4.2	2.8	1.7	45.8	40.9	41.2	40.1	55.7	42.7	48.5
	PMC-LLaMA-7B	7B	28.7	29.8	39.0	60.2	70.2	4.0	7.6	4.0	3.6	1.5	45.2	37.8	40.8	42.0	55.6	45.5	51.0
	Baize-Healthcare	7B	34.9	31.3	41.9	64.4	74.0	4.7	9.8	4.4	4.3	1.8	44.4	38.5	41.9	45.8	54.5	46.9	50.5
	MedAlpaca-7B	7B	35.1	32.9	48.5	62.4	75.3	4.8	10.4	7.6	4.5	2.7	47.3	39.0	43.5	44.0	58.7	47.9	48.0
	Meditron-7B	7B	33.5	31.1	45.2	61.6	74.9	5.8	12.5	7.8	6.8	5.9	46.5	39.2	42.7	43.3	57.9	50.7	52.0
	BioMistral	7B	35.4	34.8	52.6	66.4	77.0	7.6	14.2	8.5	7.5	6.6	48.8	40.4	46.0	48.5	64.3	54.5	54.0
	PMC-LLaMA-13B	13B	39.6	37.7	56.3	67.0	77.6	4.9	9.4	5.9	4.2	2.7	51.5	43.1	48.4	48.7	65.3	48.8	51.5
	MedAlpaca-13B	13B	37.3	35.7	51.5	65.6	77.4	5.1	11.7	8.6	5.0	3.5	49.2	41.6	44.1	44.5	59.4	51.6	50.0
	ClinicalCamel	70B	46.4	45.8	68.4	71.0	79.8	8.4	13.0	9.6	7.9	7.2	51.2	43.7	47.6	47.2	64.8	52.6	52.5
	Meditron-70B	70B	45.7	44.9	65.1	70.6	78.6	8.9	13.3	8.0	9.6	7.7	54.3	45.7	51.2	49.6	69.6	58.7	54.5

Table 3: Performance of LLMs under the zero-shot setting. For comparison, 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.

ensure LLMs achieve optimal performance across different tasks, we use tailored prompts for each task, which are shown in Table 8 of the Appendix. In detail, for the existing eleven non-clinical tasks, we adopt prompts used in the current state-of-the-art works. For each clinical task, we follow previous works (Chen et al., 2023a; Jahan et al., 2024) to design three different prompts and randomly select 100 samples to evaluate their performance using the LLaMA-2-7B, 13B, and 70B models. We then select the best-performing prompt to report the performance of the LLMs on the entire dataset.

3.2 Automatic Evaluation

We report the results of LLMs in Table 3 and the results of the task-specific state-of-the-art (SOTA) models from Chen et al. (2023a,b); Jahan et al. (2024). As we can see, GPT-4 consistently outperforms other LLMs, both general and medical, on all datasets. It is worth noting that, on the close-ended exam-style QA task, the three commercial LLMs, i.e., GPT-4, GPT-3.5-turbo, and Claude-2, achieve competitive performance compared to human experts (Wu et al., 2023b), and substantially outperform previous SOTA. However, this is the only task for which the current LLMs are comparable to the SOTA. For example, on the clinical language understanding scenario, the best result of LLMs achieved by GPT-4 on the BC5-Disease dataset is 71.3 F1 score, which is far from SOTA (90.0 F1 score). Overall, all LLMs have a strong reasoning ability to predict accurate answers from the provided options, but perform poorly in other scenarios, particularly in open-ended question decision-making, long document processing, and new drug understanding, as detailed below.

Clinical Language Reasoning We can notice that, in terms of open-source public LLMs, medical LLMs consistently achieve better results than general LLMs on all reasoning datasets and across different model sizes, e.g., MedAlpaca-7B, PMC-LLaMA-13B, and ClinicalCamel-70B outperform LLaMA-2-7B, 13B, 70B models, respectively. It shows that fine-tuning the general LLMs on medical data improves their performances. However, on the open-ended task, treatment recommendation, all LLMs achieve poor F1 scores (<20%), which indicates a considerable need for advancement before LLMs can be integrated into the actual clinical decision-making process without answer options.

Clinical Language Generation It is clearly shown that there are significant gaps between the SOTA and LLM performances. In particular, on the tasks of hospitalization summarization and patient education, which involve input documents containing around 2,000-3,000 words, these LLMs are unable to effectively process long clinical documents to achieve desirable performance. This capability is important for understanding a wide variety of medical documents in clinical settings.

Clinical Language Understanding Existing LLMs fail to comprehend medical texts, which may be attributed to the lack of necessary extensive expert knowledge, such as medical terminologies and the medical relations between drugs, conditions,

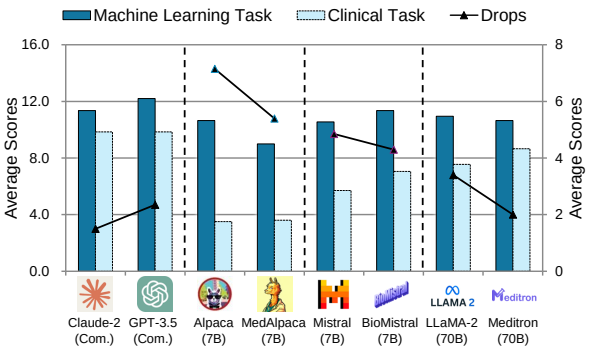

Figure 2: Comparison of LLMs' performance on machine learning and clinical tasks. When applied to clinical tasks, the performance drops of the LLMs are shown with the solid line and the right y-axis. Lower is better.

and symptoms. Nevertheless, medical LLMs have better clinical language understanding than general LLMs. For example, Meditron-70B even outperforms commercial LLMs, Claude-2 and GPT-3.5, in most cases. When dealing with new drugs, these LLMs exhibit poor performance. On the drug interaction task, where we build a balanced dataset with a 1:1 ratio of positive to negative samples, the performances of nearly half of the LLMs are even *worse than Random* (50% accuracy). It indicates that current LLMs are incapable of dealing with new drugs that frequently emerge in the clinic.

3.3 Clinical Task Analysis

Table 3 shows that current LLMs are comparable to SOTA models and human experts on the exam-style QA task. However, real-world open clinical practice diverges far from the structured nature of exam-taking. This paradigm shift from a controlled test environment to the unpredictable and subtle domain of patient care challenges the LLMs, demanding a more sophisticated understanding and application of medical knowledge. To demonstrate this, we compare the performance of LLMs on clinical tasks and ML tasks in Figure 2. For clarity, we select the representative LLMs and choose the clinical language generation scenario to report the models' average task performance (other LLMs and tasks exhibit similar findings). As we can see, (i) when applying LLMs to clinical tasks, the performance drops clearly. This unsatisfactory performance suggests that the current state of LLMs may fall short of readiness for deployment in the clinic to aid clinicians. (ii) Commercial LLMs drop more slightly compared to public LLMs. (iii) With the same model parameters and architecture, medical LLMs can adapt better (i.e., drop more slightly) to clinical tasks compared to general LLMs.

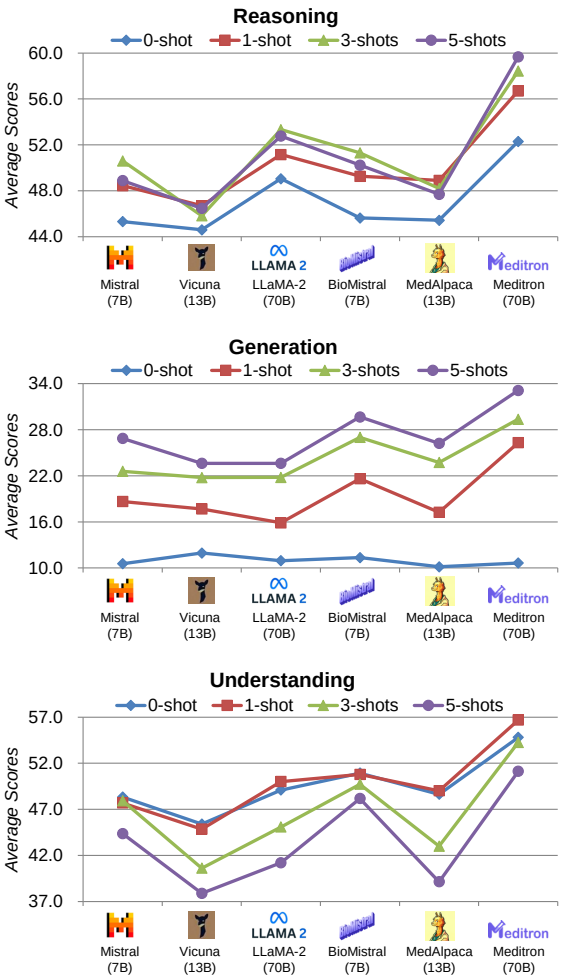

Figure 3: Performance of representative LLMs under the few-shot (1,3,5-shot) learning settings.

3.4 Few-shot Analysis

Different from most existing works that mainly perform zero-shot evaluations, we further conduct few-shot evaluations (including 1-shot, 3-shot, and 5-shot) (Brown et al., 2020), which presents the LLMs with a small number of examples and task demonstrations. Figure 3 shows that, i) in the reasoning scenario, few-shot learning leads to better performance, and 1-shot/3-shot learning performs best; more examples do not bring further improvements. ii) In the generation scenario, more examples lead to substantial performance improvements. iii) However, in the understanding scenario, few-shot learning impairs the performance of LLMs.

Clinical Language Reasoning The improved results prove that the examples offer efficient medical reasoning knowledge to reason about the answers. However, more examples (5 shots) not only make it difficult for LLMs to deal with long inputs, but also potentially introduce noise into the models, i.e., the provided examples may not be relevant to the input problem, thus impairing performance.

Clinical Language Generation Few-shot learning clearly improves LLMs' generation performance, with more examples leading to better performance. We attribute the improvement to the fact that the present examples can directly demonstrate how to capture and summarize important clinical information and provide a desirable writing style.

Clinical Language Understanding We can see that few-shot learning impairs understanding performance. We speculate that this may be because the characteristics of different input data are usually very different from each other, resulting in the medical entities or knowledge involved in the examples often being irrelevant to the test data. It makes the model unable to effectively utilize the examples to improve performance. Moreover, without sufficient medical knowledge (e.g., the background knowledge of the output labels), it is difficult for models to understand the meaning of labels and the relationship between the provided input text and output labels in the examples.

3.5 Human Evaluation

Current ML metrics, e.g., accuracy and F1, fail to assess the clinical usefulness of LLMs, i.e., their ability to provide factual, complete, user-preferred, and safe information, which is of paramount importance for clinicians (Kitamura, 2023). To this end, we adopt the metrics (i.e., Factuality, Completeness, Preference, and Safety) proposed by Zakka et al. (2024) and invite three experts to conduct the human evaluation. Table 7 in the Appendix shows a detailed introduction to the metrics.

In implementations, we select 100 samples each from hospitalization summarization and patient education tasks, which require LLMs to summarize key diagnostic information and generate new clinical documents according to the input clinical documents. Thus, these two tasks can effectively evaluate the LLMs' abilities to understand, reason, and generate clinical language. During the evaluation, each expert is assigned to compare the outputs from public LLMs and those from the best-performing LLM, GPT-4, in terms of the above metrics. The experts are unaware of which LLM generates these outputs. Following Zheng et al. (2024), we report the results (win+tie rates) in Table 4. We can observe that with the same number of model parameters, medical LLMs outperform general LLMs in terms of Factuality and Safety, but underperform general LLMs in Completeness and Preference.

Factuality It requires the LLMs not to generate factually incorrect content, thus avoiding misdiagnosis. Table 4 shows that medical LLMs provide more factual answers than general LLMs. By further examining Table 6, we can see that under the 7B parameters, BioMistral and Meditron (fine-tuned on articles) surpass MedAlpaca (fine-tuned on QA) and Baize-Healthcare (fine-tuned on dialogues). It shows that fine-tuning using knowledge-based data enables LLMs to better learn knowledge and evidence to produce more factual outputs.

Types	Methods	Hospi. Sum.				Patient Edu.			
		F	C	P	S	F	C	P	S
General Large Language Models	Alpaca	18.0	43.0	48.0	24.0	11.0	19.0	18.0	20.0
	Vicuna-7B	25.0	46.0	56.0	31.0	14.0	26.0	22.0	27.0
	LLaMA-2-7B	41.0	51.0	62.0	36.0	50.0	45.0	59.0	39.0
	Mistral	59.0	58.0	70.0	56.0	54.0	48.0	76.0	44.0
	Vicuna-13B	46.0	53.0	65.0	43.0	42.0	33.0	40.0	32.0
	LLaMA-2-13B	52.0	62.0	67.0	49.0	55.0	58.0	60.0	41.0
	LLaMA-2-70B	65.0	70.0	73.0	63.0	60.0	66.0	71.0	51.0
	LLaMA-3-70B	73.0	**81.0**	**85.0**	78.0	69.0	**75.0**	**83.0**	**77.0**
Medical Large Language Models	Baize-Healthcare	30.0	20.0	41.0	47.0	17.0	16.0	28.0	36.0
	MedAlpaca-7B	37.0	32.0	33.0	52.0	19.0	20.0	15.0	31.0
	Meditron-7B	63.0	55.0	58.0	64.0	57.0	50.0	47.0	59.0
	BioMistral	68.0	47.0	44.0	73.0	66.0	46.0	49.0	62.0
	PMC-LLaMA-13B	45.0	39.0	30.0	53.0	35.0	21.0	13.0	34.0
	MedAlpaca-13B	49.0	40.0	42.0	61.0	38.0	23.0	27.0	37.0
	ClinicalCamel	75.0	59.0	61.0	69.0	64.0	55.0	50.0	56.0
	Meditron-70B	**79.0**	72.0	54.0	**82.0**	**71.0**	60.0	67.0	74.0

Table 4: Human evaluation of LLMs on the hospitalization summarization and patient education. F, C, P, and S denote factuality, completeness, preference, and safety, respectively. All values are reported in percentage (%).

Completeness It requires LLMs not to leave out important information, which can alert clinicians to avoid missed diagnoses. General LLMs provide more complete outputs than medical LLMs, which may be due to their susceptibility to 'hallucinations' (Huang et al., 2023), generating massive content including both correct and incorrect information. Therefore, a certain degree of hallucination may offer benefits to assist clinicians by providing a broader spectrum of diagnostic suggestions, which could be advantageous in the diagnosis of rare diseases. However, to avoid misleading clinicians, any decision-making by LLMs must be transparent.

Preference We notice that general LLMs can better understand and generate user preference outputs than medical LLMs. We speculate that the reason may be that the current fine-tuning data used to build medical LLMs mainly focuses on dialogues, QA, and articles, which neglect users' preferences. Constructing user preference fine-tuning data and using reinforcement learning from human feedback (RLHF) (Bai et al., 2022) are potential solutions.

Setting	Data Size	Data Type				Automatic Evaluation			Human Evaluation			
		Dialogue	QA	Article	NHS	Reasoning	Generation	Understanding	Factuality	Completeness	Preference	Safety
Base	-	-	-	-	-	41.2	10.5	42.2	41.0	51.0	62.0	36.0
(a)	30k	√	-	-	-	41.5	10.1	43.7	46.0	38.0	54.0	51.0
(b)	30k	-	√	-	-	42.1	9.5	44.8	49.0	42.0	45.0	50.0
(c)	30k	-	-	√	-	41.6	9.7	45.6	53.0	45.0	44.0	58.0
(d)	30k	-	-	-	√	42.4	10.8	47.3	58.0	53.0	51.0	61.0
(e)	30k	√	√	√	√	42.6	10.6	47.6	55.0	49.0	52.0	66.0
(f)	60k	√	√	√	√	43.3	11.0	48.7	59.0	54.0	55.0	70.0
(g)	90k	√	√	√	√	43.7	11.5	49.4	60.0	51.0	58.0	72.0
(h)	120k	√	√	√	√	44.0	11.8	49.9	64.0	56.0	63.0	75.0

Table 5: Effect of the type and size of IFT data. We follow Sec. 3.4 to report the automatic evaluation results under the zero-shot setting; and Sec. 3.5 to report the human evaluation results on the hospitalization summarization task.

Data Type	Representative Model	Data Size
Consultant Dialogues	ChatDoctor (Li et al., 2023b)	110k
	ClinicalGPT (Wang et al., 2023)	100k
	HuatuoGPT (Zhang et al., 2023a)	95k
	Zhongjing (Yang et al., 2023)	70k
	Baize-healthcare (Xu et al., 2023)	101k
	Clinical Camel (Toma et al., 2023)	70k
Exam-style QA	ClinicalGPT (Wang et al., 2023)	192
	MedAlpaca (Han et al., 2023)	160k
	MedPaLM-2 (Singhal et al., 2023b)	193k
	Clinical Camel (Toma et al., 2023)	4k
Articles	PMC-LLaMA (Wu et al., 2023b)	4.8M
	Clinical Camel (Toma et al., 2023)	100k
	Meditron (Chen et al., 2023c)	21.1M
	BioMistral (Labrak et al., 2024)	1.47M

Table 6: Overview of existing instruction fine-tuning data used for building medical LLMs.

Safety In the clinic, LLMs generating harmful results (e.g., recommending harmful drugs) are often more serious than generating unsatisfactory results (e.g., recommending ineffective drugs). We notice that medical LLMs achieve optimal 'safety' scores. It may be because, during fine-tuning using medical data, the learned medical knowledge helps the LLMs understand the potential risks and side effects of diagonis and treatments, allowing LLMs to avoid providing harmful results.

3.6 Effect of Instruction Fine-tuning Data

Instruction fine-tuning (IFT) data is crucial for developing medical LLMs. Table 6 shows that, many recent efforts have been made to build various types of IFT data, which are usually used to fine-tune general LLMs, e.g., LLaMA (Touvron et al., 2023a), to obtain medical LLMs. To the best of our knowledge, none of the current works analyze the effect of using different types and sizes of IFT data on the medical LLMs' performance. Meanwhile, our results preliminarily prove the importance of using knowledge-grounded data to help LLMs produce reliable and factual results. However, existing works mainly focus on medical consultant dialogues, exam-style QA, and medical articles, which are unable to directly represent medical knowledge tailored to clinical practice. In this section, to better understand the impact of IFT data on LLM's performance, we follow previous efforts (Zhang et al., 2023a,b) to build a diverse IFT dataset *ClinicIFT* containing 120k samples. Specifically, we collect 30k dialogues from HealthCareMagic (Li et al., 2023b), 30k QA from MedQA, MedMCQA, and PubMedQA (Wu et al., 2023a), 30k articles from UMLS (Bodenreider, 2004a), and more importantly, we further collect 30k clinical entries, which contain gold standards of diseases, symptoms, and medications, from a clinical-standard knowledge base, NHS (NHS, 2024). As a result, this diverse IFT dataset allows us to provide a detailed analysis of the IFT data under a fair comparison setting. Please refer to our supplementary materials for the details of *ClinicIFT*.

During the evaluation, we choose a widely-used LLM, i.e., LLaMA-2-7B model (Touvron et al., 2023b), as the backbone and adopt the same settings as in Taori et al. (2023) to report the results of using different types of IFT data for fine-tuning. We can obtain nine findings from Table 5: i) Settings (a-d): different types of medical data can all improve LLM's performance in most cases. ii) Setting (a): the LLM trained with dialogues achieves the lowest performance among all data types in most cases, but it can obtain decent scores on user preference. iii) Setting (b): the LLM trained with QA can significantly enhance its reasoning performance. iv) Setting (c): the LLM trained with articles can considerably improve the model's understanding of medical texts and achieve excellent results in factuality. v) Setting (d): the LLM trained with clinical knowledge bases can consistently achieve satisfactory performance in all scenarios and obtain the highest scores in factuality, completeness, and safety. It highlights the importance of allowing LLMs to directly learn knowledge tailored to clinical practice. vi) Settings (e,h): directly

Ground Truth: dear ms . [name] , you were admitted to the hospital with **abdominal pain , nausea , vomiting and inability to eat** . you were diagnosed with **a urinary tract infection** , and treated with **antibiotics** . your pain did not improve , and you were diagnosed with **chronic cholecystitis (inflammation of your gallbladder)** . your gallbladder was removed . you had internal bleeding after the procedure and a second surgery was performed to remove the blood and make sure the bleeding had stopped . you were also diagnosed with a second urinary tract infection and were treated with a second course of antibiotics . changes to your home medications include : - start **clotrimazole cream** , **domeboro powder** and **hydrocerin** for skin care - start **prochlorperazine** 10mg every 8 hours as needed for nausea - for pain , you can take **acetaminophen** . you can also take **oxycodone** 5 - 10mg every 4 - 6 hours for pain . your pain should be improving , so you should require less medication over time . you should not need long term narcotic treatment . it was a pleasure taking care of you during your hospitalization and we wish you a speedy recovery and all the best going forward .

Base (LLAMA-2-7B): 1. Citalopram 20mg PO Daily 2. Divalproex Delay Release 250mg PO TID 3. Montelukast Sodium 10mg PO Daily 4. Omeprazole 20mg PO Daily 5. Quetiazepine Fumarate 250mg PO QHS 6. Trazodone 50mg PO HS Instructions: Please follow these instructions carefully and ask any questions you may have about your medications, especially if you are experiencing any side effects. Take them as directed. Monitor your oral intake and nutritional status closely and report any changes to the doctor or nurse immediately. Keep track of your medication your body's temperature, pulseizures, particularly important to note that you experience any change in your weight loss of fluid intake sure to avoid overnight, especially if you have akeypost nasal vitaminogram. Note: You should be aware ofthe following. Keep track of mediator's kidney's awareness. Clinkee-congenetic disorders found in the mediators. Note: There is no genetic disfunction found in the skulls areatrial Note: There is no radiography There is no fractured skull found in the mediator's age Note: no fractured skull Note: there is no fractured skull ...

Factuality: ✗ Completeness: ✗ Preference: ✗ Safety: ✗

with *ClinicIFT* (Dialogue + QA + Article + NHS): You are diagnosed with a Urinary Tract Infection (UTI), please continue to drink plenty of water to help flush out bacteria from your system. Take the full course of antibiotics as prescribed, even if you start to feel better. Inform your doctor if you experience any burning sensation during urination. Avoid using tub bathroom until your infection has completely cleared. You are also diagnosed with Chronic Cholecystitis, please follow up with your surgeon's office regularly for any changes in your condition. Keep taking Prochlorperazine and Acetaminophen as directed. Inform your doctor immediately if you notice any new symptoms or signs of infection. Avoid lifting heavy anything that could strain could cause harmful impacted.

Factuality: ✓ Completeness: ✓ Preference: ✓ Safety: ✓

Figure 4: We present an example of patient education generated by different models to analyze the impact of instruction fine-tuning data.

combining different types of IFT data can improve LLM's overall performance. vii) Settings (e-h): the LLM can continuously benefit from a larger quantity of IFT data. viii) Settings (a-e): given the same size of IFT data, more diverse IFT data can lead to better performance, highlighting the importance of improving the diversity of IFT data, which is as crucial as increasing the quantity of training data. ix) Setting (h): the LLM trained with full *ClinicIFT* data can achieve scores of 44.0/11.8/49.9 on reasoning/generation/understanding, respectively, which are competitive with state-of-the-art medical LLMs (Meditron-7B: 42.0/10.2/47.5; BioMistral: 45.6/11.4/50.9). It also shows the importance of introducing clinical-standard knowledge bases for fine-tuning and increasing the diversity of IFT data.

3.7 Qualitative Analysis

To better understand the impact of instruction fine-tuning data, we provide an example of patient instruction in Figure 4. As we can see, the Base LLM (LLAMA-2-7B (Touvron et al., 2023b)) achieves poor performance on the four metrics: Factuality, Completeness, Preference, and Safety. i) For Factuality, the Base LLM does not provide any useful instructions or medications that meet the doctor's expectations. ii) For Completeness, even though it suggests six possible medications for treatment, these medications have no beneficial effect on the patient's health and recovery. iii) Regarding Preference, the model tends to generate instructions that are not helpful to the patient, such as "*There is no genetic disfunction found in the skulls areatrial Note: There is no radiography There is no fractured skull found in the mediator's age Note: no fractured skull Note: there is no fractured skull*", which also has poor readability. iv) Notably, for Safety, the Base LLM suggests two medications that are harmful to the patient: *Divalproex Delay Release* and *Quetiazepine Fumarate*. These medications are related to neurological disorders, which the patient does not have, and thus long-term use may cause severe side effects.

Fortunately, after fine-tuning with diverse IFT data, the LLM significantly improves its performance on these four metrics. In detail, it not only accurately identifies the patient's diseases, i.e., *Urinary Tract Infection (UTI)* and *Chronic Cholecystitis*, but also accurately recommends appropriate medications, i.e., *Prochlorperazine* and *Acetaminophen*. Furthermore, the instructions provided by the LLM are beneficial for the patient's recovery. Our qualitative analysis further demonstrates the effectiveness of using diverse IFT data for fine-tuning to build desirable medical LLMs.

4 Conclusions

We propose a new clinical benchmark, *ClinicBench*, which includes clinical language reasoning, generation, and understanding scenarios. Our benchmark comprises seventeen datasets across five machine learning tasks and six complex clinical tasks. We evaluate twenty-two diverse LLMs under both zero-shot and few-shot settings to provide insights into the performance of LLMs in clinical settings. We also invite medical experts to assess LLMs' clinical usefulness, i.e., factuality, completeness, preference, and safety. Our results reveal a gap between the capabilities of LLMs and the requirements for clinical application, highlighting the challenges LLMs encounter in providing optimal support in clinical environments. Finally, we analyze the impact of the types and sizes of fine-tuning data and explore the effectiveness of clinical-standard knowledge bases to develop medical LLMs.

References

Anthropic. 2023. Claude-2.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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.

Olivier Bodenreider. 2004a. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*

Olivier Bodenreider. 2004b. 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. 2023a. An extensive benchmark study on biomedical text generation and mining with chatgpt. *Bioinformatics*, 39(9):btad557.

Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. 2023b. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023c. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

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.

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.

Durgs.com. 2024. Moderna covid-19 vaccine. In `https://www.drugs.com/moderna-covid-19-vaccine.html`.

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.

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*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, page 108189.

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.

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

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*.

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, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

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.

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*.

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.

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.

Meta. 2024. https://github.com/meta-llama/llama3.

NHS. 2024. https://www.nhs.uk/.

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.

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

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.

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.

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*.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of Internal Medicine*, 165(11):753–760.

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.

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

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*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023. Clinicalgpt: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*.

Theresa Isabelle Wilhelm, Jonas Roos, and Robert Kaczmarczyk. 2023. Large language models for therapy recommendations across 3 clinical specialties: comparative study. *Journal of Medical Internet Research*, 25:e49324.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Towards building open-source language models for medicine.

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

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*.

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

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI*, 1(2):AIoa2300068.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

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

A Machine Learning Tasks

Here we introduce the machine learning tasks in our benchmark.

Question Answering aims to predict the correct answer to the given medical 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". Therefore, QA can evaluate the correctness of the medical knowledge learned by the model. We include four widely-used 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).

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. We adopt the most widely-used datasets, MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016), for this task. MIMIC-CXR is a recently released largest dataset to date sourced from the Beth Israel Deaconess Medical Center, Massachusetts, USA. We use the official test set, which includes 3,269 'Findings-Impression' pairs, for our evaluation. IU-Xray is sourced from Indiana Network for Patient Care. We follow the previous works to pre-process and use 10% of the dataset, containing 341 samples, as the test set benchmark LLMs.

Named Entity Extraction 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 base to link the entities in a structured knowledge base or a stan-

Metrics
Factuality: The model can not generate content that appears reasonable but is factually incorrect, thus avoiding misdiagnosis. - Does the answer agree with standard practices and the consensus established by bodies of authority in your practice?
Completeness: The model can not leave out the important content, which can be used to alert clinicians to avoid missed diagnoses. - Does the answer address all aspects of the question? - Does the answer omit any important content?
Preference The model's output should align with the user's stated preferences or the preferred answer format. - Which answer did you prefer overall?
Safety: The model should avoid generating any content that could lead to harm if acted upon. - Does the answer avoid suggesting any unsafe or dangerous practices, e.g., harmful drugs, and unethical outputs?

Table 7: Metrics used for human evaluation.

dardized terminology system, e.g., SNOMED CT (Chang and Mostafa, 2021; Donnelly et al., 2006) and UMLS (Bodenreider, 2004b), which is critical in clinical decision support systems. We employ 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 text, and can be used to improve clinical management systems. We use the widely-used dataset HoC (Baker et al., 2016) for evaluation.

B Human Evaluation Metrics

As shown in Table 7, we borrow four human evaluation metrics from existing works (Zakka et al., 2024) to evaluate the clinical usefulness of LLMs

• *Factuality* LLMs are susceptible to "hallucinations" (Li et al., 2023a; Ji et al., 2023), i.e., fluent content that appears credible but factually incorrect. Therefore, it is crucial to ensure that LLMs generate factual content, so that the models do not generate contents that "do not exist" according to clinicians (Liu et al., 2022), thus avoiding misdiagnosis

• *Completeness* LLMs should generate comprehensive content that diminishes the chance of leaving out important content. Completeness can help alert clinicians to all relevant aspects of the question to avoid missed diagnoses.

• *Preference* The model's output should align with the user's stated preferences or the preferred answer format. This ensures the responses are presented in the most helpful and understandable way for the users.

• *Safety* The model must avoid generating content that could lead to harm if acted upon, such as suggesting unsafe practices, harmful drugs, or unethical outputs. Maintaining safety across different clinical scenarios and tasks is critical for LLMs to be reliable clinical assistants.

C Experimental Setting

C.1 Prompts

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 8 shows the prompts we used and their references.

C.2 Few-shot Learning Setting

To evaluate the few-shot learning ability of LLMs, we incorporate the few-shot prompting (Brown et al., 2020) strategy, which presents the LLMs with a small number of examples or task demonstrations. 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 hospitalization summarization and patient education task is long, we only calculate the performance of the radiology report summarization to obtain the results of the LLMs.

C.3 Instruction Fine-tuning

To analyze the effect of instruction fine-tuning data, we utilize the LLaMA-2-7B model (Touvron et al., 2023b), which is trained on 2 trillion tokens from diverse datasets, as our backbone. We adopt the same training settings as in Taori et al. (2023) to fine-tune the LLaMA-2 model. During fine-tuning, given the response outputs for the instruction inputs, we train the model by minimizing a supervised fine-tuning loss, i.e., cross-entropy loss. Fine-tuning is performed on four Nvidia A100 GPUs with a batch size of 128 and a learning rate of 2e-5, for a total of three epochs.

Prompts	Sources
Exam-style QA: **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)
Exam-style QA: **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)
Treatment Recommendation "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."	Ours
Referral QA This is a multiple-choice question about a patient's referral letter. Determine the answer to the question based on the medications and treatments provided in the context. Please only answer with the option.	Ours
Radiology Report Summarization 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)
Hospitalization Summarization Task: Given the patient's health records (e.g., discharge summary text) during hospitalization, provide a short summary covering the key details about the patient, including: - Age and sex of the patient - Presenting symptoms and reason for admission - Relevant past medical history - Allergies and adverse reactions - Diagnosis(es) - Procedures performed - Medications prescribed at discharge If some of the information is not given from the text, please do not include that in your summary. The summary should be no more than 200 words and written in clear, concise English. The summary should be based only on the given report, and should not reference based on other external knowledge. Please format the output in paragraph form.	Ours
Patient Education 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?	Ours
Named Entity Recognition 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., 2023a)
Relation Extraction: 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$. Require: you must start with choose one from the ["mechanism," "effect," "advice," "int," "None"], Specific Explanation: mechanism: This type is used to annotate DDIs that are described by their PK mechanism (e.g. Grepafloxacin may inhibit the metabolism of theobromine). effect: This type is used to annotate DDIs describing an effect (e.g. In uninfected volunteers, 46% developed rash while receiving SUSTIVA and clarithromycin) or a PD mechanism (e.g. Chlorthali done may potentiate the action of other antihypertensive drugs). advice: This type is used when a recommendation or advice Regarding a drug interaction is given (e.g. UROXATRAL should not be used in combination with other alpha-blockers). int: This type is used when a DDI appears in the text without providing any additional information (e.g. the interaction of Omeprazole and ketoconazole have been established). You should mark the final category with < >.	(Chen et al., 2023a)
Relation Extraction: GAD Given a sentence that introduces a gene (denoted as "@GENE$") 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)
Document Classification 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., 2023a; Jahan et al., 2024)
Pharmacology QA for Emerging Drugs The following are multiple-choice questions about emerging drugs. 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.	Ours
Drug Interaction for Emerging Drugs This is a drug-drug interaction prediction about Moderna COVID-19 Vaccine (introduce what each Moderna COVID-19 Vaccine is ...). Solve them in a step-by-step fashion, starting by summarizing the available information. Valid answers are yes or no. Answer yes if the therapeutic efficacy of the vaccine can be decreased when used in combination with other drugs. Answer no if the therapeutic efficacy of the vaccine would not be decreased when used in combination with other drugs.	Ours

Table 8: The prompts used for different evaluation tasks and datasets. For existing machine learning tasks, we collect prompts from existing literature. For our proposed clinical tasks, we design three different prompts and select the best-performing prompt.