npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01390-4

Towards evaluating and building versatile large language models for medicine

Chaoyi Wu[1,2,5], Pengcheng Qiu[1,2,5], Jinxin Liu[3], Hongfei Gu[4], Na Li[3], Ya Zhang[1,2], Yanfeng Wang[1,2] ✉ & Weidi Xie[1,2] ✉

In this study, we present **MedS-Bench**, a comprehensive benchmark to evaluate large language models (LLMs) in clinical contexts, **MedS-Bench**, spanning 11 high-level clinical tasks. We evaluate nine leading LLMs, *e.g.*, MEDITRON, Llama 3, Mistral, GPT-4, Claude-3.5, *etc*. and found that most models struggle with these complex tasks. To address these limitations, we developed **MedS-Ins**, a large-scale instruction-tuning dataset for medicine. **MedS-Ins** comprises 58 medically oriented language corpora, totaling 5M instances with 19K instructions, across 122 tasks. To demonstrate the dataset's utility, we conducted a proof-of-concept experiment by performing instruction tuning on a lightweight, open-source medical language model. The resulting model, **MMedIns-Llama 3**, significantly outperformed existing models on various clinical tasks. To promote further advancements, we have made **MedS-Ins** fully accessible and invite the research community to contribute to its expansion. Additionally, we have launched a dynamic leaderboard for **MedS-Bench**, to track the development progress of medical LLMs.

Large Language Models (LLMs) have recently achieved significant advancements across various natural language processing tasks, demonstrating remarkable capabilities in language translation, text generation, dialogue, and beyond. These developments have also extended into the medical domain, where LLMs have achieved high scores on multiple-choice question-answering (MCQA) benchmarks in healthcare, and successfully passed the UMLS examination, as noted by Singhal et al.[1,2]. Moreover, LLMs have shown expert-level performance in clinical text summarization when appropriate prompting strategies are employed[3].

Alongside these advancements, however, there has been growing criticisms and concerns regarding the application of LLMs in clinical settings, primarily due to their deficiencies in fundamental medical knowledge. For instance, LLMs have demonstrated poor comprehension of ICD codes[4], produced inaccurate predictions related to clinical procedures[5], and misinterpreted Electronic Health Record (EHR) data[6]. We posit that these polarized views on the efficacy of LLMs arise from the stringent standards required for AI deployment in clinical environments. Current benchmarks, which largely focus on multiple-choice problems[2,7,8], fail to adequately reflect the practical utility of LLMs in real-world clinical scenarios.

To address this gap, we introduce **MedS-Bench** (S for Super), a comprehensive benchmark that extends beyond multiple-choice question answering (MCQA), to include **11 advanced clinical tasks**, such as *clinical report summarization*, *treatment recommendations*, *diagnosis*, and *named entity recognition*, *among others*. This benchmark provides clinicians and researchers with a detailed understanding of where LLMs excel and where they fall short in medical tasks. Specifically, we evaluate nine mainstream models for medicine: MEDITRON[8], Mistral[9], InternLM 2[10], Llama 3[11], Qwen 2[12], Baichuan 2[13], Med42-v2[14], GPT-4[15] and Claude-3.5[16]. Our findings indicate that even the most advanced LLMs struggle with complex clinical tasks, even when utilizing few-shot prompting, underscoring the gap between high performance on MCQA benchmarks and the actual demands of clinical practice.

To advance the development of open-source medical LLMs capable of tackling a broad spectrum of clinical tasks, we take inspiration from the idea of Super-NaturalInstructions[17], and construct the first, comprehensive instruction tuning dataset for medicine, **MedS-Ins**. It aggregates **58** open-source biomedical natural language processing datasets from five text sources, including exams, clinical texts, academic papers, medical knowledge bases, and daily conversations, resulting in **5M instances with 19K instructions** across **122 clinical tasks**, each accompanied with hand-written task definitions. We performed extensive instruction tuning on open-source medical language models, and explored both zero-shot and few-shot prompting strategies. The outcome is a new medical LLM—**MMedIns-Llama 3**, for the first time, showing the effectiveness of training on diverse medical tasks through instruction tuning, enabling open-source medical

[1]Shanghai Jiao Tong University, Shanghai, China. [2]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [3]China Mobile Communications Group Co., Ltd., Beijing, China. [4]China Mobile Communications Group Shanghai Co., Ltd., Shanghai, China. [5]These authors contributed equally: Chaoyi Wu, Pengcheng Qiu. ✉e-mail: wangyanfeng622@sjtu.edu.cn; weidi@sjtu.edu.cn

LLMs to surpass leading closed-source models, including GPT-4 and Claude-3.5, across a wide range of clinical tasks.

While our final model serves primarily as an academic proof of concept, we believe that **MedS-Ins** represents an initial step toward advancing medical LLMs for real-world clinical applications, moving beyond the confines of online chat or multiple-choice question answering.

Results

In this section, we first introduce **MedS-Bench**, the benchmark employed in our study, designed to provide a comprehensive evaluation across a range of tasks critical to clinical applications. We then present detailed statistics on our instruction tuning dataset, **MedS-Ins**, which was carefully curated to cover a broad spectrum of medical language processing tasks. Finally, we provide an in-depth analysis of the evaluation results, comparing the performance of leading mainstream models with our own model, **MMedIns-Llama 3**, adapted from an open-source language model and fine-tuned on comprehensive medical instructions.

To ensure clarity in our subsequent discussion and analysis, we define key terminologies used throughout this study. For additional examples, please refer to the "Detail Tasks in MedS-Ins" section of the Supplementary.

- **Text domains:** Refers to the nature or type of the data, such as clinical texts, examination materials, academic papers, and so forth.
- **Data sources:** In contrast to the "text domains" which describe the attribute of the data, "data sources" refer to the specific origins of the data, such as MIMIC-IV or PubMed papers. Different data sources may belong to the same text domain.
- **Task categories:** These denote the broad types of language processing tasks, such as multiple-choice question answering or named entity recognition, *etc*. Tasks within the same category share a common objective.
- **Tasks:** These denote the fundamental units (leaf nodes) in our data collection pipeline, including specific tasks like outcome extraction, drug dose extraction, pathology summarization, *etc*. Each task may be defined by unique combinations of data sources, task categories, or text domains.

First of all, to evaluate the capabilities of various LLMs in clinical applications, we developed **MedS-Bench**, a comprehensive medical benchmark that extends beyond traditional multiple-choice questions. **MedS-Bench** encompasses **11 high-level clinical task categories**, derived from **28 existing datasets**, as illustrated in Fig. 1. Each dataset was reformatted into an instruction-prompted question-answering structure, complete with hand-crafted task definitions (instructions), as shown in Fig. 2a. The task categories we considered include: *Multi-choice Question Answering, Text Summarization, Information Extraction, Explanation, Rationale, Named Entity Recognition, Diagnosis, Treatment Planning, Clinical Outcome Prediction, Text Classification, Fact Verification, and Natural Language Inference*. A more detailed description of each category is provided in the "Task Category Details" section of the Supplementary.

In addition to defining these task categories, we also provide detailed statistics on the number of tokens and distinguish the required competencies for LLMs to address each task, as presented in the "Detail Tasks in

Category	Number
Diagnosis	134529
Text Summarization	41238
NER	37585
Information Extraction	16856
MCQA	13701
Clinical Outcome Prediction	4500
Treatment Planning	3535
Explanation	4486
NLI	2844
Fact Verification	2142
Text Classification	158

Item	Number
Categories	11
Source Dataset	28
Tasks	52
Samples	0.3M

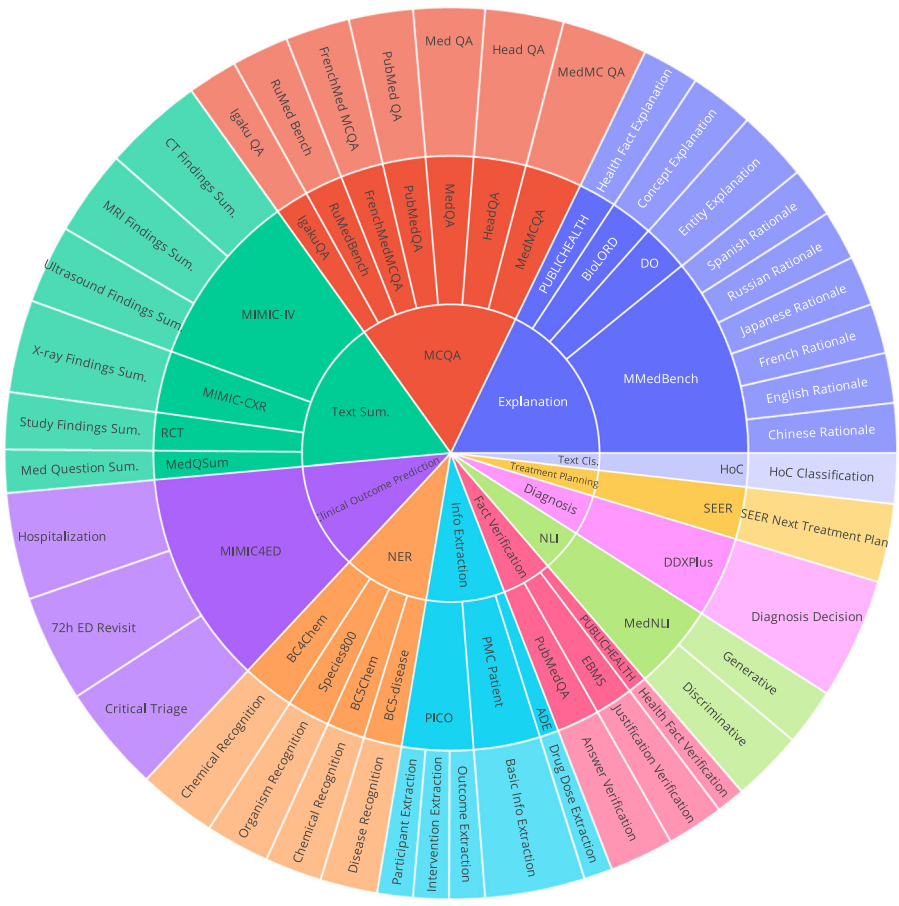

Fig. 1 | Benchmark Statistics. The hierarchical ring chart meticulously displays the data distribution within the evaluation benchmarks. The first tier categorizes the types of tasks, with the benchmarks encompassing 11 primary task categories. The second tier outlines the datasets involved, including 28 datasets in total. The third tier details the specific tasks, with the benchmarks collectively addressing 52 distinct tasks. Overall, this benchmark allows for a thorough and comprehensive evaluation of model performance across multiple dimensions.

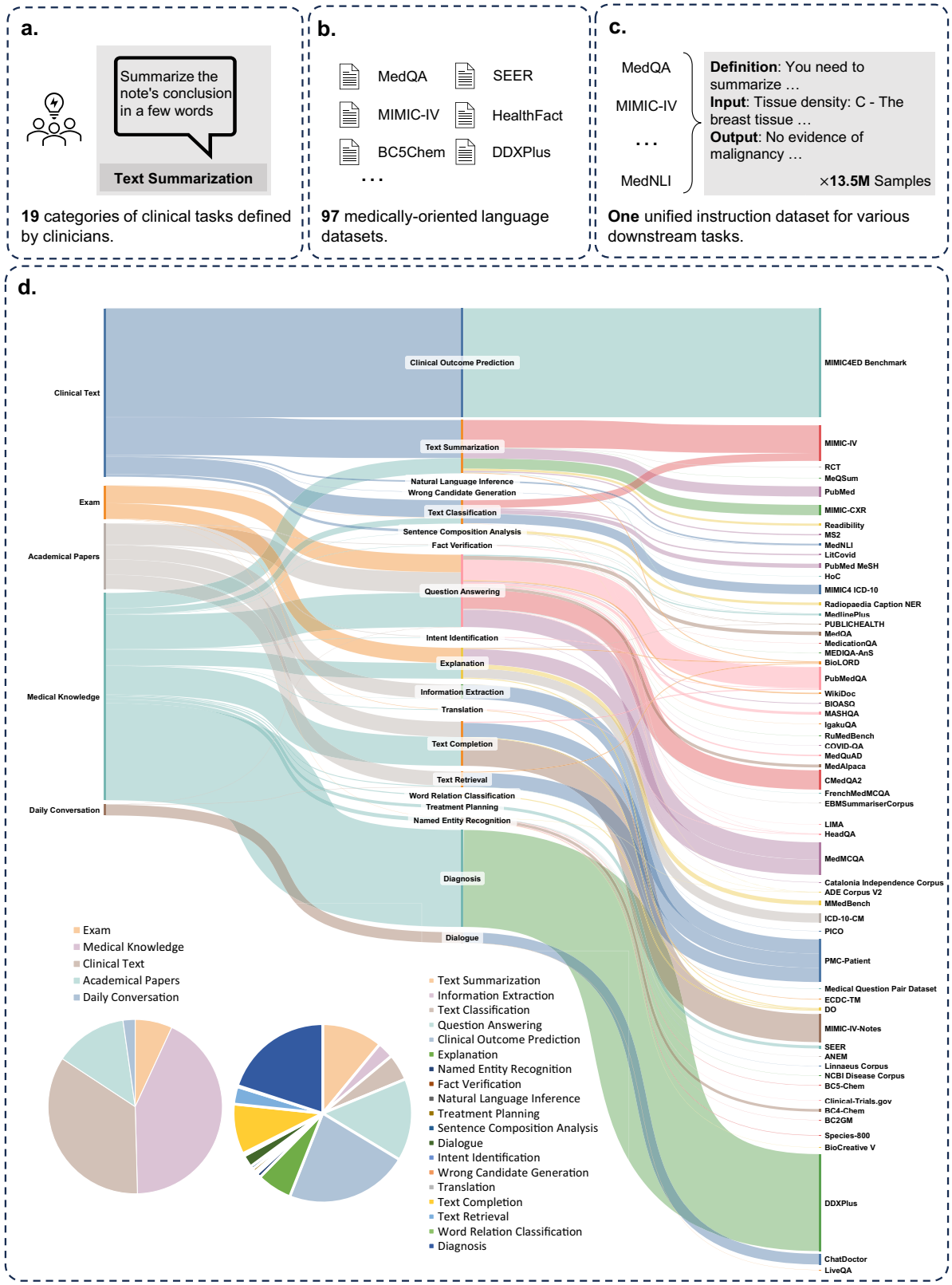

Fig. 2 | Overview of MedS-Ins. a The task collection pipeline. For each task, we add a task category along with a hand-written definition to it, resulting in a total of 19 task categories. **b** We collect the existing 58 public datasets. **c** We convert the formats of different datasets into one unified medical instruction dataset, **MedS-Ins**. **d** The final data distribution of our collected **MedS-Ins**. The Sankey diagram shows how the different text domains (left), task categories (middle), and data sources (right) contribute to the final datasets. On the left of the bottom, two pie charts show the data distributions on text domains and task categories respectively.

MedS-Ins" section of the Supplementary. Following previous work[18], we manually classified the tasks into two categories based on the skills required: (i) recalling facts from the model, and (ii) retrieving facts from the provided context. Broadly speaking, the former involves tasks that require to access knowledge encoded in the model's weights from large-scale pre-training, while the latter involves tasks that necessitate extracting information from the provided context, such as in summarization or information extraction. As shown in the "Detail Tasks in MedS-Ins" section of the Supplementary, eight of the task categories require the model to recall knowledge from the model, while the remaining three require fact retrieval from the given context.

Then, we introduce our proposed instruction dataset, **MedS-Ins**, with the data collected from 5 distinct text sources and 19 task categories, 122 distinct clinical tasks. The statistics of **MedS-Ins** are summarized in Fig. 2. Our proposed instruction tuning dataset is composed of samples drawn from five distinct **text domains**: exams, clinical texts, academic papers, medical knowledge bases, and daily conversations, introduced as follows:

- **Exams:** This category consists of data from medical examination questions across various countries. It encompasses a broad spectrum of medical knowledge, ranging from fundamental medical facts to complex clinical procedures. While the exam domain is a vital resource for understanding and assessing medical education, it is important to note that the highly standardized nature of exams often results in over-simplified cases compared to real-world clinical tasks. 7% of the tokens in our dataset are from the exams.
- **Clinical texts:** Generated during routine clinical practice, these texts support diagnostic, treatment, and preventive processes within hospitals and clinical centers. This category includes Electronic Health Records (EHRs), radiology reports, lab results, follow-up instructions, and medication recommendations, among others. These texts are indispensable for disease diagnosis and patient management, making accurate analysis and understanding crucial for the effective clinical application of LLMs. 35% of the tokens in our dataset are from clinical texts. Notably, the significant proportion of clinical texts ensures that the instruction tuning data aligns closely with clinical demands.
- **Academic papers:** This data is sourced from medical research papers, covering the latest findings and advancements in the medical research field. Given their accessibility and structured organization, extracting data from academic papers is relatively straightforward. These cases help models grasp cutting-edge medical research information, guiding them to better understand contemporary developments in medicine. 13% of the tokens in our dataset are from academic papers.
- **Medical knowledge bases:** This domain comprises well-organized and comprehensive medical knowledge, including medical encyclopedias, knowledge graphs, and glossaries of medical terms. Such data forms the backbone of medical knowledge bases, supporting both medical education and the application of LLMs in clinical practice. 43% of the tokens in our dataset are from medical knowledge.
- **Daily conversations:** This source refers to the daily consultation generated between doctors and patients, primarily sourced from online platforms and other interactive scenarios. These interactions reflect the real-life interactions between medical professionals and patients, playing a critical role in understanding patients' needs and enhancing the overall experience of medical service. 2% of the tokens in our dataset are from daily conversations.

Beyond categorizing the text domains from which the original data is sourced, the samples in **MedS-Ins** are further organized into distinct **task categories**. We have identified 19 task categories, each representing a critical capability that we believe a medical LLM should possess. By constructing this instruction-tuning dataset and fine-tuning models accordingly, we aim to equip the LLMs with the versatility needed to address a broad spectrum of medical applications.

These 19 task categories include but are not limited to, the 11 categories in the MedS-Bench benchmark. The additional categories encompass a range of linguistic and analytical tasks essential for comprehensive medical language processing, including *Intent Identification, Translation, Word Relation Classification, Text Retrieval, Sentence Composition Analysis, Wrong Candidate Generation, Dialogue*, and *Text Completion* and the MCQA category is extended to general *Question Answering*, which also includes free-text answering cases. The diversity of task categories ranging from common question answering and dialogue to various downstream clinical tasks—guarantees a comprehensive understanding of the potential medical applications. A detailed description of each category is provided in the "Task Category Details" section of the Supplementary.

After introducing all our proposed datasets, we will analyze different LLMs on various tasks accordingly. For each task type, we start by discussing the performance of various existing LLMs, followed by a comparison with our final model, **MMedIns-Llama 3**. All results presented here were obtained using a 3-shot prompting strategy (more details in the "Evaluation Settings" section in Supplementary), except for MCQA tasks, where we used a zero-shot prompting setting to align with previous studies[7,8,19]. As our comparisons include proprietary models like GPT-4 and Claude 3.5, which incur usage costs, we randomly sampled around 1500 test cases per benchmark to manage the cost constraints. The task description and specific sampling numbers are detailed in the section "Task Category Details" in Supplementary. For simplicity, the percentage mark (%) is omitted in all the following tables and results analysis.

Multilingual Multiple-choice Question-answering

Here, we present evaluation results on the widely used multiple-choice Question-answering (MCQA) benchmarks, as shown in Table 1. Some numbers are directly incorporated from our previous studies[8,13,14,19,20]. On these multi-choice question-answering datasets, existing proprietary LLMs have demonstrated very high accuracies, for example, on MedQA, GPT-4 can achieve 85.8, which is almost comparable to human experts, and Llama 3 can also pass the exam with 60.9 scores. Similarly, in languages other than English, LLMs also demonstrate superior results in multiple-choice accuracy on MMedBench[19]. The results indicate that as multi-choice questions have been extensively considered in existing research, different LLMs may have been specifically optimized for such tasks, resulting in high performance. It is therefore essential to build up a more comprehensive benchmark, to further push the development of LLMs towards clinical applications. Our proposed model, MMedIns-Llama 3, although not primarily trained on multi-choice questions, still shows notable improvement, achieving an average accuracy of 63.9 across different benchmarks, significantly surpassing GPT-3.5.

Text Summarization

As shown by Table 2, the performance of text summarization is reported as 'BLEU/ROUGE' scores, on multiple report types across various modalities, including X-ray, CT, MRI, ultrasound, and other medical questions. Among the models, closed-source LLMs, such as GPT-4 and Claude-3.5, perform better than all the open-source ones, achieving an average of 24.46/25.66 and 26.29/27.36, respectively. Among open-source models, Mistral achieves the best results, with BLEU/ROUGE scores of 24.48/24.90. Llama 3 follows closely, with scores of 22.20/23.08. Our model (MMedIns-Llama 3), trained on the medical-specific instruction dataset (**MedS-Ins**), significantly outperforms the others, achieving average scores of 46.82/48.38.

Information Extraction

The performance of information extraction is summarized in Table 3. InternLM 2 shows exceptionally good performance in this task with an average score of 79.11. For example, In the PICO tasks, InternLM 2 leads in both Intervention and Outcome Extraction, with scores of 74.42 and 69.77, respectively. Closed-source models such as GPT-4 and Claude-3.5 outperform all other open-source counterparts, with average scores of 76.92 and 79.41, respectively. Analysis of individual benchmark components reveals that most LLMs perform better at extracting less medically complex information, such as basic patient details, compared to more specialized

Table 1 | Results on medical multiple-choice question answering, as reported with Accuracy score

Method[1]	Size	MedQA	MedMCQA	PubMedQA	MMedBench					Avg.
					ZH	JA	FR	RU	ES	
Close-source Models										
GPT-3.5	–	57.7	**72.7**	53.8	52.3	34.6	32.5	66.4	66.1	54.5
GPT-4	–	**85.8**	72.3	75.2	75.1	**72.9**	**56.6**	**83.6**	**85.7**	**75.3**
Open-source Models										
MEDITRON	7B	47.9	59.2	74.4	61.9	40.2	35.1	67.6	53.3	54.9
InternLM 2	7B	–	–	–	77.6	47.7	41.0	68.4	59.6	-
Mistral	7B	50.8	48.2	75.4	71.1	44.7	48.7	74.2	63.9	49.1
Llama 3	8B	60.9	50.7	73.0	78.2	48.2	50.8	71.5	64.2	62.2
Qwen 1.5	7B	48.9	50.2	67.8	–	–	–	–	–	–
Med42-v2	8B	62.8	62.8	75.8	–	–	–	–	–	–
Baichuan 2	7B	32.7	41.7	–	–	–	–	–	–	–
MMedIns-Llama 3	8B	63.6	57.1	**78.2**	**78.6**	54.3	46.0	72.3	61.2	63.9

Bolding represents the best results. Notably, the results except MMedIns-Llama 3 in this table are all borrowed from other works.
[1]The results for GPT-3.5, GPT-4, MEDITRON, InternLM, Mistral, and Llama 3 are borrowed from MMedLM[19]. Med42-v2 and Baichuan 2 are borrowed from their original papers[13,14]. Qwen 1.5 is borrowed from Open-Medical-LLM-Leaderboard[20]. Notably, since we do not find the reported scores for the latest Qwen 2, the earlier Qwen 1.5 (Qwen/Qwen1.5-7B) is instead reported to represent this LLM family.

Table 2 | Results on text summarization, as reported with BLEU and ROUGE scores, formatted as 'BLEU/ROUGE'

Method	Size	MedQSum	RCT-Text	MIMIC-CXR	MIMIC-IV			Avg.
		Med Question	Study Con.	X-ray	Ultrasound	CT	MRI	
Close-source Models								
GPT-4	–	25.06/27.30	34.32/31.09	27.26/29.71	11.17/14.53	23.97/29.52	25.76/32.06	24.46/25.66
Claude-3.5	–	21.14/25.06	41.02/36.16	27.76/29.93	15.24/18.28	21.98/26.38	26.43/31.05	26.29/27.36
Open-source Models								
MEDITRON	7B	15.64/23.14	4.00/16.44	5.21/16.50	3.75/6.07	16.30/23.93	20.11/27.98	7.15/15.54
InternLM 2	7B	15.69/21.63	14.48/15.16	11.83/13.41	13.48/20.96	20.88/27.82	23.43/31.40	13.87/17.79
Mistral	7B	23.49/26.03	27.24/26.13	22.09/24.71	25.09/22.72	27.60/30.77	29.87/31.81	24.48/24.90
Llama 3	8B	22.45/25.08	15.38/14.60	32.92/32.64	18.06/20.00	24.47/29.35	24.82/30.50	22.20/23.08
Qwen 2	7B	16.41/19.29	25.37/24.20	14.64/14.87	17.20/18.15	18.38/22.36	22.80/26.67	18.41/19.13
Med42-v2	8B	12.34/18.61	36.45/32.99	16.82/16.87	15.07/16.21	23.13/24.80	27.44/28.78	20.17/21.17
Baichuan 2	7B	12.74/17.30	23.32/24.09	13.55/12.94	14.91/16.31	19.68/20.92	25.21/26.25	16.13/17.66
MMedIns-Llama 3	8B	**54.16/56.95**	**57.82/55.60**	**54.91/57.64**	**20.40/23.32**	**42.18/46.46**	**40.53/43.38**	**46.82/48.38**

"Con." denotes conclusions. Bolding represents the best results.

medical data like outcomes and interventions. For instance, in extracting basic information from PMC Patients, most LLMs score above 90, with Claude-3.5 achieving the highest score of 99.07. In contrast, performance on Clinical Outcome Extraction tasks within PICO is relatively poor. Our proposed model, MMedIns-Llama 3, demonstrates the best overall performance, achieving an average score of 83.77, surpassing InternLM 2 by 4.66 points. Notably, in the PICO tasks, MMedIns-Llama 3 excels in Participant Extraction, scoring 83.72, which exceeds the second-best model by 11.63 points.

Concept Explanation

We conduct evaluations on medical concept explanation and reported the BLEU-1 and ROUGE-1 scores across all relevant datasets and models. In Table 4, we evaluate the model on medical concept explanation, GPT-4 performs well on this task, achieving average scores of 19.37/21.58. In contrast, MEDITRON and Qwen 2 achieve relatively lower scores, with averages of 8.51/18.90 and 9.20/12.67. We hypothesize that MEDITRON's lower performance may be due to its training corpus, which is primarily focused on academic papers and guidelines, making it less effective at explaining basic medical concepts. Similarly, Qwen 2's performance appears limited by its lack of familiarity with medical terminology. While Qwen 2 performs better on the Health Fact Explanation task, which is less domain-specific, its performance drops significantly on more specialized tasks, such as DO entity explanation and BioLord concept explanation, where models are required to explain medical terms. Our final model, MMedIns-Llama 3, significantly outperforms the other ones across all concept explanation tasks, particularly in Health Fact Explanation (30.50/28.53) and BioLORD Concept Explanation (38.12/43.90), and achieving the highest average scores of 34.43/37.47. Following MMedIns-Llama 3, GPT-4 also showed strong performance, with GPT-4 scoring 19.37/21.58.

Answer Explanations (Rationale)

In Table 5, we evaluate the complex rationale, *i.e.*, explaining the answer and comparing the reasoning abilities of various models using the **MMedBench**[19] dataset across six languages. Among the models tested, the closed-source model Claude-3.5 exhibited the strongest performance, with average scores of 46.26/36.97, demonstrating consistently high scores across all languages, particularly in French and Spanish. This superior performance

Table 3 | Results on information extraction, as reported with Accuracy score

Method	Size	PICO			ADE Drug Dose Ext.	PMC patient Basic Info. Ext.	Avg.
		Participant Ext.	Intervention Ext.	Outcome Ext.			
Close-source Models							
GPT-4	–	67.44	62.79	65.12	91.30	97.93	76.92
Claude-3.5	–	65.12	76.74	60.47	95.65	**99.07**	79.41
Open-source Models							
MEDITRON	7B	72.09	46.51	51.16	**95.65**	72.20	67.52
InternLM 2	7B	72.09	74.42	**69.77**	**95.65**	83.60	79.11
Mistral	7B	60.47	65.12	48.84	91.30	85.20	70.18
Llama 3	8B	58.14	79.07	58.14	69.57	95.93	72.17
Qwen 2	7B	58.14	67.44	41.86	73.91	95.93	67.46
Med42-v2	8B	55.81	60.47	60.47	91.30	95.67	72.74
Baichuan 2	7B	48.84	34.88	16.28	69.57	73.33	48.58
MMedIns-Llama 3	8B	**83.72**	**79.07**	62.79	**95.65**	97.60	**83.77**

"Ext." denotes extraction and "Info." denotes information. Bolding represents the best results.

Table 4 | Results on medical concept explanation, as reported with 'BLEU/ROUGE' scores

Method	Size	Health Fact Exp.	Do Entity Exp.	BioLORD Concept Exp.	Avg.
Close-source Models					
GPT-4	–	18.63/20.80	19.14/21.14	20.33/22.80	19.37/21.58
Claude-3.5	–	14.96/18.48	8.75/13.28	13.95/18.49	12.56/16.75
Open-source Models					
MEDITRON	7B	6.09/8.65	7.68/25.39	11.76/22.66	8.51/18.90
InternLM 2	7B	22.36/27.01	5.28/10.39	6.95/13.62	11.53/17.01
Mistral	7B	18.11/21.31	9.21/14.11	13.27/16.68	13.53/17.37
Llama 3	8B	16.79/20.32	14.88/18.84	8.87/14.61	13.51/17.92
Qwen 2	7B	14.94/17.45	5.87/9.73	6.81/10.83	9.20/12.67
Med42-v2	8B	18.15/21.21	13.31/17.13	12.26/15.64	14.57/18.00
Baichuan 2	7B	18.04/20.56	9.75/13.12	10.99/13.62	12.93/15.77
MMedIns-Llama 3	8B	**30.50/28.53**	**34.66/39.99**	**38.12/43.90**	**34.43/37.47**

"Exp." denotes Explanation. Bolding represents the best results.

Table 5 | Results on rationale, as reported with 'BLEU/ROUGE' scores

Method	Size	MMedBench						Avg.
		Chinese	English	French	Japanese	Russian	Spanish	
Close-source Models								
Claude-3.5	-	44.64/34.63	47.07/38.67	**48.93/41.23**	49.22/**39.15**	**38.90/28.17**	**48.80/39.99**	46.26/**36.97**
Open-source Models								
MEDITRON	7B	20.39/21.79	38.42/31.24	34.43/29.33	18.89/24.98	24.32/16.77	37.64/31.01	29.01/25.86
InternLM 2	7B	35.23/30.77	44.12/37.39	36.10/33.65	29.13/33.15	27.43/20.99	41.87/36.30	35.65/32.04
Mistral	7B	35.53/28.91	47.20/37.88	39.53/35.64	29.16/28.96	32.15/23.99	45.27/38.33	38.14/32.28
Llama 3	8B	28.51/23.30	44.10/39.26	24.92/22.24	13.46/15.04	31.16/22.85	32.37/27.70	29.09/25.06
Qwen 2	7B	41.53/29.89	43.67/34.22	30.39/27.72	46.78/33.54	24.89/22.15	40.09/36.38	37.89/30.65
Med42-v2	8B	19.42/17.55	47.22/39.45	32.01/26.71	10.85/11.52	26.87/20.35	32.00/24.58	28.06/23.36
Baichuan 2	7B	32.09/26.70	39.52/32.09	17.74/17.57	14.63/13.52	18.38/15.06	31.85/28.12	25.70/22.18
MMedIns-Llama 3	8B	**50.27/34.01**	**49.08/38.19**	46.93/38.73	**51.74**/35.19	35.27/23.81	48.15/37.35	**46.90**/34.54

Note here, we do not include the results for GPT-4 since the original evaluation sets were generated with it, which may bring unfair comparison bias. Bolding represents the best results.

Table 6 | Results on NER tasks, as reported with F1-Score scores

Method	Size	BC4Chem Chemical Rec.	BC5Chem	BC5Disease Disease Rec.	Species800 Organism Rec.	Avg.
Close-source Models						
GPT-4	-	54.84	67.62	53.20	62.43	59.52
Claude-3.5	-	22.98	40.77	24.05	14.45	25.56
Open-source Models						
MEDITRON	7B	1.98	4.11	1.33	0.40	1.96
InternLM 2	7B	41.21	41.51	37.11	62.93	45.69
Mistral	7B	15.56	32.09	12.17	6.31	16.53
Llama 3	8B	19.45	37.83	25.30	11.90	23.62
Qwen 2	7B	29.17	39.60	20.70	52.72	35.55
Med42-v2	8B	23.51	36.52	26.95	11.92	24.73
Baichuan 2	7B	11.96	18.68	11.47	34.56	19.17
MMedIns-Llama 3	8B	**90.78**	**91.25**	**54.26**	**80.87**	**79.29**

'Rec.' is short for 'recognition'. Bolding represents the best results.

may be attributed to the similarity of this task to chain-of-thought reasoning, a capability that has been specifically enhanced in many general-purpose LLMs. Among open-source models, Mistral and InternLM 2 demonstrated comparable performance, with average scores of 38.14/32.28 and 35.65/32.04, respectively. It is important to note that GPT-4 was excluded from this evaluation, because the rationale component of the MMedBench dataset was primarily constructed using GPT-4 outputs, which could introduce bias and bring unfair comparisons. Consistent with our observations in concept explanation, our final model, MMedIns-Llama 3, demonstrated the best overall performance with average BLEU-1/ROUGE-1 scores of 46.90/34.54 across all languages, notably, achieving 51.74/35.19 in Japanese reasoning tasks, 49.08/38.19 in English, and 46.93/38.73 in French, respectively. This superior performance is likely due to the fact that our base language model (MMed-Llama 3) was initially developed to be multilingual[19]. Consequently, even though our instruction tuning did not explicitly target multilingual data, the final model outperforms others across multiple languages.

Named Entity Recognition (NER)

As shown in Table 6, among these models GPT-4 is the only one that consistently demonstrates robust performance across Named Entity Recognition (NER) tasks, achieving an average F1-Score of 59.52. It excels particularly in the BC5Chem Chemical Recognition task with a score of 67.62. InternLM 2 shows the best performance among all the open-source models, with an average F1-Score of 45.69. Qwen 2 and Med42-v2 also show solid performance, with averages of 35.55 and 24.73, respectively. Llama 3 and Mistral, with average F1-Score of 23.62 and 16.53, respectively, exhibit moderate performance. MEDITRON, not optimized for NER tasks, shows limited effectiveness in this area. Our model, **MMedIns-Llama 3**, significantly outperforms all other models, achieving an average F1-Score of 79.29. It excels in the BC4Chem and BC5Chem Chemical Recognition tasks, with F1 scores of 90.78 and 91.25, respectively. Furthermore, MMedIns-Llama 3 leads in the BC5Disease Disease Recognition task with an F1-Score of 54.26 and in the Species800 Organism Recognition task with 80.87, demonstrating superior capability in handling complex NER tasks across various biomedical domains.

Diagnosis, Treatment Planning, and Clinical Outcome Prediction

We evaluate the performance on tasks involving diagnosis, treatment planning, and clinical outcome prediction, using the DDXPlus benchmark for Diagnosis, the SEER benchmark for Treatment Planning, and the MIMIC4ED benchmark for Clinical Outcome Prediction. The results, presented in Table 7, are measured in terms of accuracy. Here, the use of accuracy as a metric is appropriate in this generative prediction problem, as each of these datasets simplifies the original problem into a closed-set choice. Specifically, DDXPlus utilizes a predefined list of diseases, from which models must select one based on the provided patient context. In SEER, treatment recommendations are categorized into eight high-level categories, while in MIMIC4ED, the final clinical outcome decisions are binary, with options of either True or False. Overall, the open-source LLMs underperform the closed-source LLMs in these tasks. In the SEER treatment planning task, InternLM 2 and MEDITRON achieved relatively high accuracy scores of 62.33 and 68.27, respectively, among open-source models, but this is still far away from GPT-4's 84.73 and Claude-3.5's 92.93. For the DDXPlus diagnosis task, the open-source exhibited similar, relatively low scores around 35.00, with even medical-specific models like MEDITRON and med42-v2 performing poorly. This may be due to the task's complexity, which differs significantly from multiple-choice QA, and the models' lack of specialized training. Notably, in clinical outcome prediction, both Baichuan 2 and Llama 3 struggled to predict critical triage and the 72-hour ED revisit binary indicator, failing to provide meaningful predictions. This may be because the task lies outside their training distribution, and the models often fail to follow the provided three-shot format, resulting in extremely low scores. Closed-source models such as GPT-4 and Claude-3.5 demonstrate significantly better performance. Claude-3.5, for instance, achieves a 92.93 accuracy score in treatment planning and GPT-4 attains 84.73. They also demonstrate better performance in diagnosis, highlighting the considerable gap between open-source and closed-source LLMs. Despite these results, the scores remain insufficient for reliable clinical use. In contrast, **MMedIns-Llama 3** demonstrate superior accuracy in clinical decision support tasks, with a 98.47 accuracy score in treatment planning, 97.53 in diagnosis, and an average accuracy of 63.35 (mean on the scores of Hospitalization, 72h ED Revisit, and Critical Triage) in clinical outcome prediction.

Text Classification

In Table 7, we present the evaluation on the Hallmarks of Cancer (HoC) multi-label classification task, and report macro-Precision, macro-Recall, and macro-F1 scores. For this task, all candidate labels are input into the language model as a list, and the model is asked to select its preferred answers, allowing for multiple selections. The metrics are then calculated based on these model selections. GPT-4 and Claude-3.5 perform well on this task, with GPT-4 achieving a macro-F1 score of 68.06 and Claude-3.5 slightly worse at 66.74. Both models show strong recall capabilities, particularly GPT-4, which achieves a macro-Recall of 80.23, underscoring its proficiency in identifying relevant labels. Among open-source models,

Table 7 | Results on treatment planning, diagnosis clinical outcome prediction, and text classification results

Method	Size	SEER	DDXPlus	MIMIC4ED			HoC Classification		
				Hospitalization	72h ED Revisit	Critical Triage	Precision	Recall	F1
Close-source Models									
GPT-4	-	84.73	58.13	61.20	**58.07**	60.13	61.07	80.23	68.06
Claude-3.5	-	92.93	60.24	65.80	57.91	**68.53**	58.43	79.84	66.74
Open-source Models									
MEDITRON	7B	68.27	29.53	56.27	48.47	45.67	19.61	34.61	23.70
InternLM 2	7B	62.33	35.20	58.80	55.13	52.80	20.65	82.24	31.09
Mistral	7B	38.93	34.80	56.27	48.47	45.67	40.39	64.11	48.73
Llama 3	8B	56.07	33.73	39.07	9.27	8.80	32.40	52.03	38.37
Qwen 2	7B	22.27	34.07	57.60	56.67	53.53	37.78	53.81	40.29
Med42-v2	8B	43.87	34.13	57.87	55.20	46.60	49.95	53.12	47.87
Baichuan 2	7B	16.80	34.13	22.73	8.07	2.13	38.54	20.28	23.76
MMedIns-Llama 3	8B	**98.47**	**97.53**	**74.20**	52.73	63.13	**89.59**	**85.58**	**86.66**

The first 3 tasks are reported with Accuracy scores, and text classification is reported with Precision, Recall, and F1 scores. Bolding represents the best results.

Table 8 | Results on fact verification and NLI results, as reported with both accuracy and BLUE/ROUGE scores

Method	Size	PubMedQA	PUBLICHEALTH	EMBS	MedNLI textual entailment	
		Answer Ver.	Health Fact Ver.	Justification Ver.	Discriminative Task	Generative Task
Close-source Models						
GPT-4	-	66.15	78.60	16.28/16.27	86.63	**27.09/23.71**
Claude-3.5	-	11.54	62.04	14.77/16.45	82.14	17.80/20.02
Open-source Models						
MEDITRON	7B	25.23	32.66	11.58/15.78	60.83	4.42/14.08
InternLM 2	7B	**99.23**	76.94	8.75/14.69	84.67	15.84/19.01
Mistral	7B	57.38	69.78	15.98/16.43	71.59	13.03/15.47
Llama 3	8B	94.77	63.89	**16.52/16.49**	63.85	21.31/22.75
Qwen 2	7B	18.00	58.25	12.52/14.00	82.00	14.26/16.21
Med42-v2	8B	73.23	78.54	15.63/15.86	77.57	12.24/15.29
Baichuan 2	7B	79.38	47.98	14.97/15.99	53.94	14.99/17.27
MMedIns-Llama 3	8B	97.08	**79.55**	12.71/14.65	**86.71**	23.52/25.17

'Ver.' denotes 'verification'. Bolding represents the best results.

Med42-v2, the latest medical LLM that underwent comprehensive supervised fine-tuning and preference optimization, performs well among the open-source models, achieving a macro-F1 score of 47.87. Llama 3 and Qwen 2 show moderate performance, with macro-F1 scores of 38.37 and 40.29, respectively. These models, especially InternLM 2, exhibit high recall but struggle with precision, resulting in lower F1 scores. Baichuan and MEDITRON rank the lowest in this task, with macro-F1 scores of 23.76 and 23.70, respectively. Our **MMedIns-Llama 3** clearly outperforms all other models, achieving the highest scores across all metrics, with a macro-Precision of 89.59, a macro-Recall of 85.58, and a macro-F1 score of 86.66. These results highlight MMedIns-Llama 3's superior ability to accurately classify and recall multiple labels, making it the most effective model for this complex task.

Fact Verification

In Table 8, we evaluate the models on fact verification tasks. For PubMedQA Answer Verification and HealthFact Verification, the LLMs are required to select a single answer from a list of provided candidates, with accuracy serving as the evaluation metric. In contrast, for EBMS Justification Verification, where the task involves generating free-form text, performance is assessed using BLEU and ROUGE scores. InternLM 2 achieves the highest accuracy on PubMedQA Answer Verification with scores of 99.23. On PUBLICHEALTH, Med42-v2 achieves 79.54, which is the best among all the open-source models, just behind GPT-4 with 78.60. In the EBMS benchmark, Llama 3 and GPT-4 show comparable performance, with average BLEU/ROUGE score of 16.52/16.49 and 16.28/16.27. **MMedIns-Llama 3** continues surpassing existing models, achieving the highest accuracy score as InternLM 2, excelling in PubMedQA Answer Verification and HealthFact Verification while in EMBS, MMedIns-Llama 3 slightly falls behind the GPT-4 and Llama 3 with 12.71/14.65 in BLEU and ROUGE, which we treat as future work for further improvement.

Natural Language Inference (NLI)

Table 8 shows the evaluation on medical Natural Language Inference (NLI) using the MedNLI textual entailment dataset. The results are measured with accuracy for the discriminative tasks (selecting the right answer from a list of candidates) and BLEU/ROUGE metrics for the generative tasks (generating free-form text answers). InternLM 2 achieves the highest scores among the open-source LLMs, scoring 84.67. For the closed-source LLMs, GPT-4 and Claude-3.5 all show relatively high scores, with 86.63 and 82.14 accuracy

scores respectively. Qwen 2 and Med42-v2 also show second-best performance of 82.00 and 77.57 among the open-source LLMs In the generative task, Llama 3 demonstrates the highest consistency with the reference ground truth, achieving scores of 21.31 for BLEU and 22.75 for ROUGE among the open-source models. Similarly, GPT-4 also performs well in the generative task format, resulting in 27.09/23.71 scores while Claude-3.5 is not ideal in this task. **MMedIns-Llama 3** achieves the highest accuracy in the discriminative task, scoring 86.71, comparable with GPT-4. MMedIns-Llama 3 also excels in the generative task, with BLEU/ROUGE scores of 23.52/25.17, outperforming other models except the GPT-4.

Run Time Analysis

Beyond the task-wise performance, we also compare the inference cost of different models. The results are shown in the "Run Time Analysis" section in Supplementary. Generally, the run-time differences between various LLM series (like Mistral vs. Llama 3) are not significant. Thus, in real clinical applications, the performance we believe is the main fact that clinicians need to consider.

Discussion

Overall, this paper makes several key contributions:

Firstly, we construct a comprehensive evaluation benchmark - MedS-Bench. The development of medical LLMs has largely relied on benchmarks focused on multiple-choice question answering (MCQA). However, this narrow evaluation framework risks overlooking the broader capabilities required for LLMs in various clinical scenarios. In this work, we introduce **MedS-Bench**, a comprehensive benchmark designed to assess the performance of both closed-source and open-source LLMs across diverse clinical tasks, including those that require fact recall from the model or reasoning from given context. Our results reveal that while existing LLMs perform exceptionally well on MCQA benchmarks, they struggle to align with the actual clinical practice, particularly in tasks such as treatment planning and explanation. This finding underscores the need for further efforts to develop medical LLMs that are better suited to a wider range of clinical and medical scenarios beyond MCQA.

Secondly, we introduce a new comprehensive instruction tuning dataset - MedS-Ins. We have developed **MedS-Ins**, a novel medical instruction tuning dataset, by extensively sourcing data from existing BioNLP datasets and converting these samples into a unified format, with semi-automated prompting strategies. Previous efforts have focused primarily on constructing question-answer pairs from daily conversations, exams, or academic papers, often neglecting the texts generated from real clinical practice. In contrast, MedS-Ins integrates a broader range of medical text sources, encompassing five primary text domains and 19 task categories, as illustrated in Fig. 2d. This systematic analysis on data composition is crucial for aligning LLMs with the diverse queries encountered in clinical practice.

Thirdly, we present a strong large language model for medicine - MMedIns-Llama 3. On the model front, we demonstrate that by conducting instruction tuning on **MedS-Ins**, we can significantly enhance the alignment of open-source medical LLMs with clinical demands. Our final model, **MMedIns-Llama 3**, serves as a proof-of-concept, featuring a medium-scale architecture with 8 billion parameters, has exhibited a deep understanding of various clinical tasks and adapts flexibly to multiple medical scenarios through zero-shot or few-shot instruction prompts, without the need for further task-specific training. As evidenced by the results, our model outperforms existing LLMs, including GPT-4, Claude-3.5, across a range of medical benchmarks, covering different text sources.

Lastly, we need highlight the limitations of our paper and the potential improvements in future work.

First, **MedS-Bench** currently covers only 11 clinical tasks, which does not fully encompass the complexity of all clinical scenarios. Additionally, while we evaluated nine mainstream LLMs, some models remain absent from our analysis. To address these limitations, we plan to release an open leaderboard for medical LLMs alongside this paper. This initiative aims to encourage contributions from the community to continually expand and refine comprehensive benchmarks for medical LLMs. Specifically, this will involve updating the test set to better reflect real clinical demands and include a broader range of medical LLMs. By incorporating more task categories from diverse text sources into the evaluation process, we hope to gain a deeper understanding of the ongoing advancements in LLMs within the medical field.

Second, although **MedS-Ins** now encompasses the widest range of medical tasks available, it remains incomplete, and certain practical medical scenarios may be missing. To address this, we have made all our collected data and resources available as open-source on GitHub. We encourage contributions from the broader AI4medicine community to help maintain and dynamically expand this instruction tuning dataset, similar to efforts for Super-NaturalInstructions in the general domain[21]. Detailed guidelines are provided on our GitHub page, and we will acknowledge every contributor involved in updating the dataset. The current limited number of tasks may explain why we have not yet observed the models exhibiting emergent abilities to generalize to unseen clinical tasks, a capability seen in LLMs trained on thousands of diverse tasks in the general domain[17,22].

Third, we plan to incorporate more languages into **MedS-Bench** and **MedS-Ins** to support the development of more robust multilingual LLMs for medicine. Multilingual language models have seen substantial development in general domains, evidenced by advancements in models[23–25], training datasets[26,27], and evaluation benchmarks[28–30]. Despite this, the multilingual capabilities of biomedical language models, particularly those dealing with diverse healthcare data from various regions, remain underexplored. Recent efforts, such as BioMistral[31], Apollo[32], and MMedLM[19], have begun to address this gap by developing multilingual medical large language models (LLMs). However, their evaluation or instruction tuning progress still mainly focuses on multiple-choice question-answering formats. This may be attributed to the lack of well-established, comprehensive benchmarks or a task-wise taxonomy in the medical field, even in English, which complicates the creation of multilingual evaluation benchmarks via translate-and-filter strategies[31]. Thus, our MedS-Bench and MedS-Ins, although currently primarily in English, can offer an exemplary taxonomy-wise framework for expansion into multilingual contexts. Expanding to include a broader range of languages would be a promising future direction, ensuring that the latest advancements in healthcare AI can benefit a wider and more diverse range of regions equitably. We leave this as a crucial potential future direction. For now, we just combine a few existing multilingual benchmarks, for example, multiple-choice question-answering, and translation.

Forth, our model has not yet undergone extensive clinical validation. We aim to collaborate with the community to develop higher-quality instruction-tuning datasets that can better reflect real clinical needs. Furthermore, we are considering to further align the model safety with human preference. With these refinements, we plan to conduct clinical validation in real-world deployments to assess its practical effectiveness in future work. Beyond the model performance, more importantly, we have to emphasize that while our benchmark is more comprehensive and clinically relevant than previous MCQA benchmarks, it cannot replace the final stage of evaluating LLMs in actual clinical settings to ensure their safety. Instead, our benchmark is expected to serve as an experimental arena for assessing the performance of different LLMs, offering a more accurate reflection of their clinical capabilities, and serving as a crucial preliminary step before costly real-world evaluations, thus significantly reducing the expenses associated with assessing a model's true clinical effectiveness.

Finally, all our code, data, and evaluation pipelines are open-sourced. We hope this work will inspire the medical LLM community to focus more on aligning these models with real-world clinical applications.

Methods

In this section, we will first describe the **data constructing procedure** for **MedS-Ins**, as shown in Fig. 3a. In order to organize the different tasks, we assign a domain tag and category tag to each task, the former denotes the

domain covered by the instructions, while the category tag denotes the applicable task. We start by filtering the medical-related sentence in natural instruction datasets, followed by prompting specific BioNLP into free-text response formats.

Filtering Natural Instructions

We start by filtering medical-related tasks from the 1616 tasks collected in Super-NaturalInstructions[21]. As this work focuses more on different natural language processing tasks in general-purpose domains, the granularity of classification is relatively coarse for the medical domain. We first extract all the instructions in "Healthcare" and "Medicine" categories, subsequently, we manually added more detailed granularity to the domain labels for them, while the task category remains unchanged.

In addition, we found that many of the organized instruction fine-tuning datasets in the generic domain also cover some healthcare-related data, such as LIMA[33] and ShareGPT[34]. To filter out the medical part of these data, we used InsTag[35] to classify the domain of each instruction at a coarse granularity. Specifically, InsTag is an LLM, specialized for tagging different instruction samples. Given an instruction query, it will analyze which domain and task it belongs to. Finally, by filtering instruction datasets in the general domain, we collect 37 tasks, for a total of 75373 samples.

Prompting Existing BioNLP Datasets

In the literature, there exist many excellent datasets on text analysis in clinical scenarios. However, as most datasets are collected for different purposes, like classification or text completion, they can not be directly used for training large language models. Here, we convert these existing former medical NLP tasks into a format that can be used for training generative models, naturally adding them into instruction tuning.

Specifically, we use MIMIC-IV-Note as an example, which provides high-quality structured reports with both findings and impressions, they are used as a proxy task for text summarization, where impressions act as an abstract summary of the findings. We first manually write prompts to define the task, for example, *"Given the detailed finding of Ultrasound imaging diagnostics, summarize the note's conclusion in a few words."*. Considering the diversity for instruction tuning, we ask 5 individuals to independently describe a certain task with 3 different prompts. This results in 15 free-text prompts for each task, with similar semantic meanings but as varied as possible in wording and format. Then, inspired by the Self-Instruct[36], we use these manually written instructions as seed prompts and asked GPT-4[15] to rewrite more task instructions based on the following prompt:

> Rewrite the following instruction definition directly. You can change the wording, but keep the meaning the same. Output the rewritten definition directly without any additional information.

Finally, for each task, we will describe it with 7 key elements as shown at the bottom of Fig. 3a, *i.e.*, *{"Categories", "Domains" "Definitions", "Input Language", "Output Language", "Instruction Language" and "Instances"}*, where "Definition" consists of the manually written or GPT-4 enhanced instruction to describe the tasks, "Input Language", "Output Language", and "Instruction Language" respectively describe the languages, such as English or Chinese, used in the corresponding components of a specific instance of this task. "Categories" and "Domains" describe what text domains and categories the task belongs to. Finally, in "Instances", different training or evaluation instances with Input and Output contents are stored.

Through the above procedure, we prompt an extra 85 tasks into a unified free-form question-answering format, combined with the filtered data, resulting in a totaling 5M instances with 19K instructions, covering 122 tasks, termed as **MedS-Ins** (the detailed 122 task information can be found in the "Detailed Tasks in MedS-Ins" section of the Supplementary, which has shown to significantly improve the LLMs on clinical tasks.

After preparing the related data, we will further detail the **training procedure**, as shown in Fig. 3b. We take the same approach as our previous work[7,19], which have shown that further auto-regressive training on medical-elated corpus can inject medical knowledge into the models, thus allowing them to perform better in different downstream tasks. We start from a multilingual LLMs base model (MMed-Llama 3[19]), and further train it with comprehensive instructions from **MedS-Ins**.

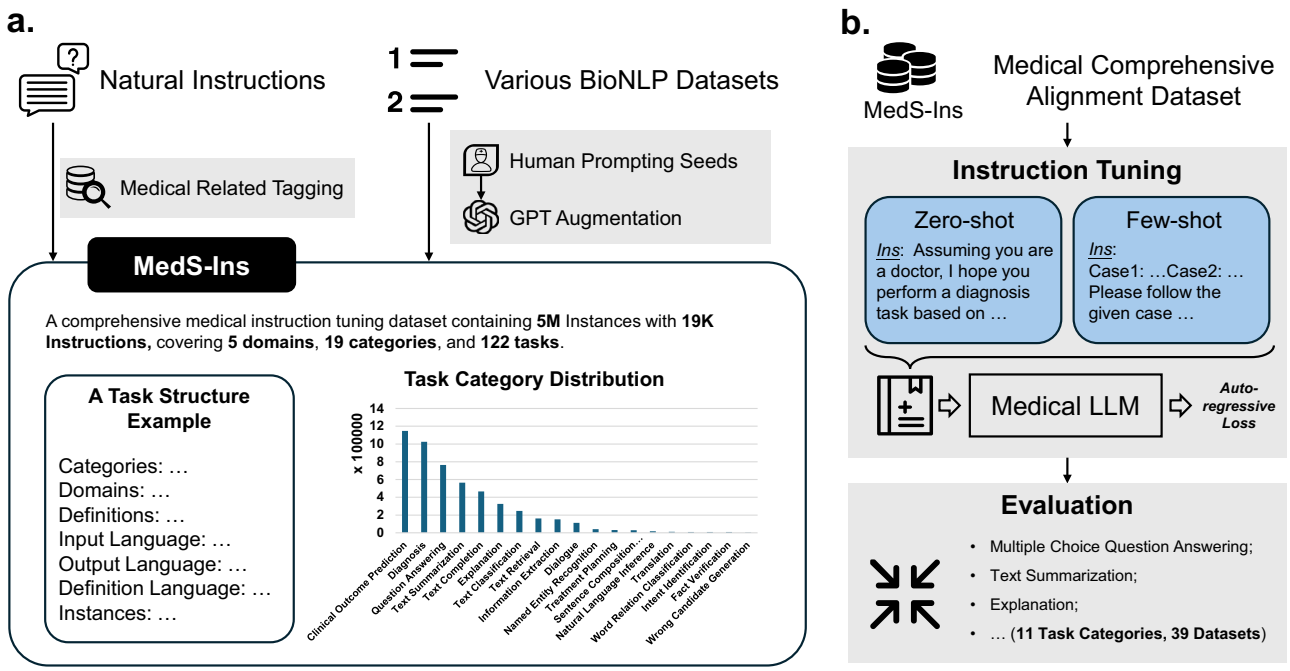

Fig. 3 | The pipeline of our method. a The data collection pipeline. We mainly collect data by filtering natural instructions and prompting well-organized BioNLP datasets. **b** The training and evaluation pipeline for our model leveraging the collected MedS-Ins. We leverage the instruction tuning training method to combine different datasets and evaluate the final model on multiple benchmarks comprehensively. The icons in the figure are from Microsoft free icon basis.

Instruction Tuning

Given the base model, trained on a large-scale medical corpus with auto-regressive prediction, we further fine-tune it to better follow human instructions or prompts. Considering an input sequence with an instruction I and a context C, and an output sequence O, the model is trained to maximize the probability:

$$P(O|C,I) = \prod_{t=1}^{|O|} P(o_t|o_1, o_2, \ldots, o_{t-1}, C, I; \theta) \quad (1)$$

Similarly, the loss function used in instruction tuning is cross-entropy loss and can be calculated as follows:

$$\text{Loss} = -\sum_{t=1}^{|O|} \log P(o_t|o_1, o_2, \ldots, o_{t-1}, C, I; \theta) \quad (2)$$

The key insight here is to construct diverse instructions, that enables the model to robustly output the preferred answers. Here, we mainly consider two types of instructions, namely, zero-shot and few-shot prompting:

- **Zero-shot Prompting**. Here, the I contains some semantic task descriptions as hints, and the model is therefore asked to directly answer the questions based on its internal model knowledge. In our collected **MedS-Ins**, the "Definition" contents for each task can be naturally used as the zero-shot instruction input. Due to the coverage of a wide range of different medical task definitions, the model is expected to learn the semantic understanding of various task descriptions. The input template is as follows:
  {INSTRUCTION}
  Input: {INPUT}
- **Few-shot Prompting**. Here, the I contains the few-shot examples, that allow the model to learn the input-output mapping on the fly. We simply obtain such instructions by randomly sampling other cases from the training set of the same task, and organizing them using a straightforward template as follows:
  Case1: Input: {CASE1_INPUT}, Output: {CASE1_OUTPUT}
  …
  CaseN: Input: {CASEN_INPUT}, Output: {CASEN_OUTPUT}
  {INSTRUCTION}
  Please learn from the few-shot cases to see what content you have to output.
  Input: {INPUT} Notably, considering some extremely long context clinical tasks the few-shot examples may exceed the max length of the context window. In this case, we adopt basic left-truncation to prioritize the last part of the case-related content or output format part in the few-shot examples.

Implementation Details

We conduct all our experiments using `PyTorch` framework and `Transformers` python package. Specifically, we set the maximum length to 2048, and pad the sequence to the longest case with padding tokens in a batch. We employ the Fully Sharded Data Parallel (FSDP) implemented with `Transformers.trainer` function to save the memory cost per GPU. We also adopt BF16 as default training precision and gradient checkpointing[37] techniques to optimize memory usage. We use a global batch size of 128 and a learning rate of 1e-5. We choose the medical-knowledge-enhanced model MMed-Llama 3 in our previous work as the foundation model. We further train the model by supervised fine-tuning on Meds-Ins for 5 Epoch with 32 Ascend910B for 383.5 hours.

At last, we will talk about our **evaluation** details. First, we provide details for the baseline large language models (LLMs). Note that, we evaluate all models in few-shot settings, as we observe that open-source models struggle to complete zero-shot evaluation. Specifically, **three** example cases are given to the model, the detailed prompting strategy and model versions can be found in the "Evaluation Settings" section in Supplementary.

The first category includes the powerful closed-source LLMs, known for their robust performance in the general domain. We evaluate these models on various medical-specific tasks:

- **GPT-4**[15], developed by OpenAI, stands for one of the most sophisticated LLMs to date. It is renowned for its strong capabilities in language processing in general domains, including medical applications.
- **Claude-3.5**[16], developed by Anthropic, is a frontier AI language model designed to be secure, trustworthy, and reliable. It exhibits advanced reasoning capabilities that enable it to perform complex cognitive tasks effectively. We adopt the Claude-3.5-Sonnet for comparison, which is claimed as the best model among the Claude family.

The second category comprises the mainstream open-source LLMs:

- **Llama 3**[11], developed by Meta AI, is one of the most notable open-source LLMs globally. As part of the LLaMA series, it is designed for high performance in natural language processing tasks, with enhancements over its predecessors in accuracy and contextual understanding. In this study, considering our model is an 8B scale LLM, for fair comparison, we adopt its 8B version as well.
- **Mistral**[9], developed by Mistral AI, is an innovative open-source LLM that claims superiority over Llama 2 13B across all evaluated benchmarks. For a fair comparison against other LLMs, we consider its 7B version.
- **Internlm 2**[10], developed by Shanghai AI Laboratory, stands out as a leading open-source multilingual LLM, showcasing exceptional performance, particularly in English and Chinese. In this paper We adopt the 7B version of Internlm 2.
- **Qwen 2**[12], developed by Alibaba, is a series of advanced large language and multi-modal models, ranging from 0.5 to 72 billion parameters, designed for high performance in a variety of tasks. In this paper we adopt the 7B version.
- **Baichuan 2**[13] is a series of large-scale multilingual language models with 7 billion and 13 billion parameters, trained from scratch on 2.6 trillion tokens. Baichuan 2 excels particularly in specialized domains such as medicine and law. Here, we adopt the 7B version.

The third category of models we choose is the open-sourced medical LLMs, which have been further trained on medical data.

- **MEDITRON**[8] is a large-scale medical LLM, further pre-trained on Llama 2. It leverages 21.1M medical papers, guidelines for further pre-training, and supervised finetuning on different MCQA datasets with context and chain-of-thought prompt styles. Similarly, we consider its 7B version.
- **Med42-v2**[14] is a suite of clinical LLMs built on the Llama 3 architecture and fine-tuned with specialized clinical data. These models are designed to address the limitations of generic LLMs in healthcare settings by effectively responding to clinical queries, which typical models avoid due to safety concerns. Similarly, we consider its 8B version.

Then, we delineate the metrics employed across various tasks and categories within our evaluation benchmark.

Accuracy

For tasks requiring the model to select a single correct answer from multiple choices, we employ 'accuracy' as a direct metric. This metric is applied to tasks *MedQA, MedMCQA, and MMedBench* in Multilingual Multiple-choice Question-answering; *participant, intervention, and outcome extraction in PICO, drug dose extraction in ADE, and patient information extraction in PMC-patient* for Information Extraction. It is also used in *SEER* for Treatment Planning, *DDXPlus* for Diagnosis, *MIMIC4ED* for Clinical Outcome Prediction, *PubMedQA and PUBLICHEALTH Verification* for Fact Verification, as well as *MedNLI textual entailment discriminative tasks* for NLI.

Precision, Recall, F1 Score

For tasks where the model is required to select multiple correct answers, we utilize Precision, Recall, and the F1 Score. These metrics are relevant for-*BC4Chem and BC5Chem for chemical recognition, BC5Disease for disease recognition, Species800 for organism recognition* in Named Entity Recognition (NER), and *HoC* in Classification.

BLEU, ROUGE

For tasks necessitating the generation of free-form text, which are inherently more challenging to evaluate, we utilize BLEU and ROUGE metrics to assess the similarity between the generated text and the ground truth. Specifically, we use BLEU-1 and ROUGE-1 by default if no other statements in this paper. These tasks include *MedQSum, RCT-Text, MIMIC-CXR, MIMIC-IV* for Text Summarization; *EBMS* for Fact Verification; *PUBLICHEALTH Explanation, Do, BioLORD and MMedBench* for Concept Explanation / Rationale; along with *generative tasks in textual entailment in MedNLI* for NLI.

Data availability

MedS-Ins is available at https://huggingface.co/datasets/Henrychur/MedS-Ins, MedS-Bench is available at https://huggingface.co/datasets/Henrychur/MedS-Bench. For datasets without redistribution licenses, we provide corresponding download links and datapreprocessing scripts at https://github.com/MAGIC-AI4Med/MedS-Ins/tree/main/data_preparing.

Code availability

Source codes of this paper is released in https://github.com/MAGIC-AI4Med/MedS-Inswith CC BY-SA license. Model weights of MMedIns-Llama 3 can be found in https://huggingface.co/Henrychur/MMedS-Llama-3-8Bwith the Llama 3 community license.

Received: 24 September 2024; Accepted: 12 December 2024;
Published online: 27 January 2025

References

1. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
2. Singhal, K. et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617* (2023).
3. Van Veen, D. et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nat. Med.* **30**, 1134–1142 (2024).
4. Soroush, A. et al. Large language models are poor medical coders—benchmarking of medical code querying. *NEJM AI* **1**, AIdbp2300040 (2024).
5. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine* 1–10 (2024).
6. Fleming, S. L. et al. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 22021–22030 (2024).
7. Wu, C. et al. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association* ocae045 (2024).
8. Chen, Z. et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).
9. Jiang, A. Q. et al. Mistral 7b. *arXiv preprint arXiv:2310.06825* (2023).
10. Cai, Z. et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297* (2024).
11. Touvron, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
12. Yang, A. et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671* (2024).
13. Yang, A. et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305* (2023).
14. Christophe, C., Kanithi, P. K., Raha, T., Khan, S. & Pimentel, M. A. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142* (2024).
15. Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y. & Radev, D. Evaluating GPT-4 and ChatGPT on Japanese medical licensing examinations (2023).
16. Anthropic Team. Introducing the next generation of claudehttps://www.anthropic.com/news/claude-3-family (2024). Accessed on March 4, 2024.
17. Wang, Y. et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 5085-5109).
18. Jin, T. et al. The cost of down-scaling language models: Fact recall deteriorates before in-context learning. *arXiv preprint arXiv:2310.04680* (2023).
19. Qiu, P. et al. Towards building multilingual language model for medicine. *Nature Communications* 15.1 (2024): 8384.
20. Pal, A., Minervini, P., Motzfeldt, A. G. & Alex, B. openlifescienceai/open_medical_llm_leaderboardhttps://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard (2024). Accessed on November 15, 2024.
21. Wang, Y. et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022* (2022).
22. Longpre, S. et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning* (pp. 22631-22648). PMLR.
23. Le Scao, T. et al. Bloom: A 176b-parameter open-access multilingual language model (2023).
24. Lai, V. D. et al. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 318–327).
25. Lu, Y., Zhu, W., Li, L., Qiao, Y. & Yuan, F. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975* (2024).
26. Nguyen, T. et al. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400* (2023).
27. Crawl, C. Common crawl maintains a free, open repository of web crawl data that can be used by anyone.https://commoncrawl.org/ (Accessed on Apr. 2024).
28. Tom, K. et al. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *WMT23-Eighth Conference on Machine Translation*, 198–216 (2023).
29. Ahuja, K. et al. Mega: Multilingual evaluation of generative ai. In *The 2023 Conference on Empirical Methods in Natural Language Processing.*
30. Zhang, W., Aljunied, M., Gao, C., Chia, Y. K. & Bing, L. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Adv. Neural Inf. Process. Syst.* **36**, 5484–5505 (2023).
31. Labrak, Y. et al. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373* (2024).
32. Wang, X. et al. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640* (2024).
33. Zhou, C. et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* **36** (2024).
34. Li, Y., Dong, B., Lin, C. & Guerin, F. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6342–6353).

35. Lu, K. et al. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations* (2023).
36. Wang, Y. et al. Self-instruct: Aligning language models with self-generated instructions. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
37. Chen, T., Xu, B., Zhang, C. & Guestrin, C. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174* (2016).

Acknowledgements

This work is supported by Science and Technology Commission of Shanghai Municipality (No. 22511106101, No. 18DZ2270700, No. 21DZ1100100), 111 plan (No. BP0719010), State Key Laboratory of UHD Video and Audio Production and Presentation, National Key R&D Program of China (No. 2022ZD0160702).

Author contributions

All listed authors clearly meet the ICMJE 4 criteria. C.W. and P.Q. contribute equally to this work. Y.W. and W.X. are the corresponding authors. Specifically, C.W., P.Q, J.L., H.G., N.L., Y.Z., Y.W., and W.X. all make contributions to the conception or design of the work, and C.W., P.Q. further perform acquisition, analysis, or interpretation of data for the work. In writing, C.W. and P.Q. draft the work. J.L., H.G., N.L., Y.Z., Y.W., and W.X. review it critically for important intellectual content. All authors approve of the version to be published and agree to be accountable for all aspects of the work to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01390-4.

Correspondence and requests for materials should be addressed to Yanfeng Wang or Weidi Xie.