

# Me-LLaMA: Medical Foundation Large Language Models for Comprehensive Text Analysis and Beyond

**Qianqian Xie**

Yale University

**Qingyu Chen**

Yale University

**Aokun Chen**

University of Florida

**Cheng Peng**

University of Florida

**Yan Hu**

University of Texas Health Science, Center at Houston

**Fongci Lin**

Yale University

**Xueqing Peng**

Yale University

**Jimin Huang**

Yale University

**Jeffrey Zhang**

Yale University

**Vipina Keloth**

Yale University

**Xinyu Zhou**

Yale University

**Lingfei Qian**

Yale University

**Huan He**

Yale University

**Dennis Shung**

Yale University

**Lucila Ohno-Machado**

Yale University

**Yonghui Wu**

University of Florida

**Hua Xu**

[hua.xu@yale.edu](mailto:hua.xu@yale.edu)

Yale University

**Jiang Bian**

University of Florida


---

**Article**

**Keywords:**

**Posted Date:** December 18th, 2024

**DOI:** <https://doi.org/10.21203/rs.3.rs-5456223/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

Recent advancements in large language models (LLMs) like ChatGPT and LLaMA have shown significant potential in medical applications, but their effectiveness is limited by a lack of specialized medical knowledge due to general-domain training. In this study, we developed Me-LLaMA, a new family of open-source medical LLMs that uniquely integrate extensive domain-specific knowledge with robust instruction-following capabilities. Me-LLaMA comprises foundation models (Me-LLaMA 13B and 70B) and their chat-enhanced versions, developed through comprehensive continual pretraining and instruction tuning of LLaMA2 models using both biomedical literature and clinical notes. Me-LLaMA utilized the largest and most comprehensive medical data, including 129B pre-training tokens and 214K instruction tuning samples from diverse biomedical and clinical data sources. Training the 70B models required substantial computational resources, exceeding 100,000 A100 GPU hours. We applied Me-LLaMA to six medical text analysis tasks and evaluated its performance on 12 benchmark datasets. To further assess Me-LLaMA's potential clinical utility, we evaluated its performance on complex clinical case diagnosis compared with other commercial LLMs, using both automatic and human evaluations. Me-LLaMA models outperform LLaMA, and other existing open-source medical LLMs in both zero-shot and supervised learning settings for most text analysis tasks. With task-specific instruction tuning, Me-LLaMA models also surpass leading commercial LLMs, outperforming ChatGPT on 7 out of 8 datasets and GPT-4 on 5 out of 8 datasets. Moreover, Me-LLaMA's performance is comparable to ChatGPT and GPT-4 for diagnosing complex clinical cases. Our findings underscore combining domain-specific continual pretraining with instruction tuning is essential for developing effective domain-specific large language models in healthcare, significantly enhancing performance across diverse medical text analysis tasks and applications. By publicly releasing our models and resources under appropriate user agreements, we aim to foster innovation and facilitate advancements in medical AI, benefiting researchers and practitioners within the community.

## INTRODUCTION

Large language models (LLMs) have shown great potential in improving medical applications such as clinical documentation, diagnostic accuracy, and patient care management.<sup>1,2,3</sup> However, general-domain LLMs often lack specialized medical knowledge because they are primarily trained on non-medical datasets<sup>4</sup>, limiting their effectiveness in healthcare settings. Although commercial LLMs, such as ChatGPT<sup>5</sup> and GPT-4,<sup>6</sup> offer advanced capabilities, their closed-source nature restricts the flexible customization and accessibility required for medical use. This limitation has spurred the research towards developing open-source LLMs such as LLaMA;<sup>7-8</sup> Yet, these models still fall short due to their general-domain training.<sup>9-10</sup>

To address these challenges, researchers have explored strategies to develop domain specific LLMs for the medical domain. Instruction fine-tuning of general-domain models, as seen in MedAlpaca,<sup>3</sup> ChatDoctor,<sup>12</sup> and AlpaCare,<sup>13</sup> attempts to enhance medical capabilities but is limited by the base models' lack of specialized knowledge; instruction fine-tuning alone cannot compensate for this deficiency. Training models from scratch using medical corpora, exemplified by GatorTronGPT,<sup>14</sup> overcomes this limitation but demands substantial computational resources and time. A more cost-effective alternative is continual pretraining, enabling models to acquire specialized medical knowledge while leveraging existing model architectures; notable examples include PMC-LLaMA,<sup>15</sup> Meditron<sup>16</sup> and Clinical LLaMA.<sup>17</sup>

**Table 1.** The comparison of Me-LLaMA models and existing open source medical LLMs.

Model	Backbone	Model size	Biomedical literature	Clinical notes	Continual pre-training (# of tokens)	Instruction tuning (# of instructions)	Evaluation tasks	Release date
MedAlpaca	LLaMA	7/13B	✓	✗	-	160K	QA	04/14/2023
ChatDoctor	LLaMA2	7B	✓	✗	-	100K	QA	05/24/2023
AlpaCare	LLaMA	7/13B	✓	✗	-	52K	QA, Summarization	10/23/2023
Clinical LLaMA	LLaMA	7B	✗	✓	-	-	Classification	07/06/2023
Meditron	LLaMA2	7/70B	✓	✗	48B	-	QA	11/27/2023
PMC-LLaMA	LLaMA	7/13B	✓	✗	79B	514K	QA	04/27/2023
<b>Me-LLaMA</b>	LLaMA2	13/70B	✓	✓	129B	214K	QA, NER, RE, Classification, Summarization, NLI, Medical Diagnosis	06/05/2024

Despite these advances, existing LLMs of continual pretraining in the medical domain exhibit notable limitations: (1) Although both domain knowledge and instruction-following capabilities are crucial, only PMC-LLaMA<sup>18</sup> has combined continual pretraining with instruction fine-tuning, revealing a gap in leveraging the synergy between these two aspects. (2) Only one model (Clinical LLaMA) used clinical notes from electronic health records, which is crucial for real-world clinical applications as it provides context-specific information from direct patient care. None of the existing models used both biomedical literature and clinical notes, which is one of the goals of this project. (3) Due to the limited medical datasets utilized for model development, these models still lack essential domain knowledge, which hampers their effectiveness. By combining biomedical literature and clinical notes, we generated the largest biomedical pre-training dataset (129B tokens), compared to the previous efforts (i.e., 79B tokens in PMC-LLaMA as the highest, see Table 1). (4) Evaluations have predominantly centered on medical question-answering (QA) tasks, lacking comprehensive assessments on the generalizability of those foundation models across diverse medical tasks.

To overcome these limitations, we present Me-LLaMA, a novel family of open-source medical large language models that uniquely integrate extensive domain-specific knowledge with robust instruction-following capabilities. Me-LLaMA comprises foundation models (Me-LLaMA 13B and 70B) and their chat-enhanced versions, developed through comprehensive continual pretraining and instruction tuning of LLaMA2 models. Leveraging the largest and most diverse medical dataset to date—combining 129 billion pretraining tokens and 214,000 instruction samples from scientific literature, clinical guidelines, and electronic health record clinical notes—Me-LLaMA excels across a wide spectrum of medical text analysis and real-world clinical tasks. Unlike prior studies, we conduct the most extensive evaluation to date, covering six critical tasks—question answering, relation extraction, named entity recognition, text classification, text summarization, and natural language inference—across twelve datasets from both biomedical and clinical domains. Our results demonstrate that Me-LLaMA not only surpasses existing open-source medical LLMs in both zero-shot and supervised settings but also, with task-specific instruction tuning, outperforms leading commercial LLMs such as ChatGPT on seven out of eight datasets and GPT-4 on five out of eight datasets. Furthermore, to evaluate Me-LLaMA’s potential clinical utility, we assessed the models on complex clinical case diagnosis tasks, comparing their performance with other commercial LLMs using both automatic and human evaluations. Our findings indicate that Me-LLaMA’s performance is comparable to that of ChatGPT and GPT-4, despite their substantially larger model sizes.

Our findings underscore the importance of combining domain-specific continual pretraining with instruction tuning to develop effective large language models for the medical domain. Recognizing the significant resources required, we have publicly released our Me-LLaMA models on PhysioNet under appropriate Data Use Agreements (DUAs) to lower barriers and foster innovation within the medical AI community. Alongside the models, we provide benchmarks and evaluation scripts on GitHub to facilitate further development. We anticipate that these contributions will benefit researchers and practitioners alike, advancing this critical field toward more effective and accessible medical AI applications.

## METHODS

We utilized LLaMA2 as the backbone model and developed Me-LLaMA through the process of continual pre-training and instruction tuning of LLaMA2, using 129B tokens and 214K instruction tuning samples from general, biomedical, and clinical domains. Figure 1

shows an overview of our study.

## Continual Pre-Training Data

To effectively adapt backbone LLaMA2 models for the medical domain through continual pre-training, we developed a mixed continual pre-training dataset, comprised of biomedical literature, clinical notes, and general domain data. It integrates over 3 million full biomedical articles from PubMed Central and over 15 million paper abstracts from PubMed, sourced from the Pile dataset.<sup>14</sup> To incorporate real-world clinical scenarios and reasoning, we included de-identified free-text clinical notes from MIMIC-III,<sup>15</sup> MIMIC-IV,<sup>16</sup> and MIMIC-CXR.<sup>17</sup> Moreover, to avoid the model forgetting acquired general knowledge, we incorporated a subset from the RedPajama<sup>18</sup> dataset, a replication of LLaMA2's pre-training data. The dataset was structured with a 15:1:4 ratio of biomedical, clinical, to general domain data and contains a total of 129 billion tokens, making it the largest pre-training dataset in the medical domain currently available.

## Medical Instruction Tuning Data

To enhance our model's ability to follow instructions and generalize across diverse medical tasks, we further developed a novel medical instruction tuning dataset with 214,595 high-quality samples from a wide array of data sources. This dataset stands out from those used in existing medical LLMs due to its comprehensive coverage of both biomedical and clinical domains. Our data sources included biomedical literature, clinical notes, clinical guidelines, wikidoc, knowledge graphs, and general domain data, as shown in Table 2. The diverse tasks aim to refine the model's ability to process and respond to medical information accurately and contextually. Detailed prompts for each data and the data example are shown in Appendix 0.1, Table A.1.

Table 2  
The overall instruction tuning dataset.

Task	Type	Source	Size	Copy right
General	Conversation	Alpaca <sup>19</sup>	20,000	CC-BY-NC 4.0
		Dolly <sup>20</sup>		CC-BY-SA-3.0
		ShareGPT <sup>21</sup>		Apache-2.0
Biomedical	Conversation	HealthCareMagic <sup>12</sup>	20,000	Reserved by HealthCareMagic and Icliniq
		Icliniq <sup>12</sup>		
	Instructions	MedInstruct <sup>13</sup>	52,000	CC BY-NC 4.0
	Question Answering	Medical Flash Cards <sup>3</sup>	34,000	No commercialized use
		MEDIQA <sup>22</sup>	2,220	CC BY 4.0
		MedicationQA <sup>23</sup>	690	CC BY 4.0
		LiveQA <sup>24</sup>	634	CC BY 4.0
		WikiDocPatient <sup>3</sup>	5,490	CC BY-SA 4.0
		GuidelineQA	2,000	Common Crawl (other)
		Summarization	PubMed Central	10,000
	Next Sentence Generation	PubMed Central	20,000	CC BY
	Key words prediction	PubMed Central	10,000	CC BY
	Causal Relation Detection	PubMed <sup>25</sup>	2,450	CC BY
	Relation Extraction	UMLS knowledge graph <sup>2</sup>	10,000	Openrail
Clinical	QA, summarization, classification, mortality prediction	MIMIC-III, <sup>15</sup> MIMIC-IV <sup>16</sup>	30,000	PhysioNet credentialed health data use agreement 1.5.0

## Training Details

As shown in Fig. 3, we developed the Me-LLaMA 13B and 70B base models by continual pre-training the LLaMA2 13B and 70B models. These base models were then instruction-tuned to create the Me-LLaMA-13B-chat and Me-LLaMA-70B-chat models.

### Me-LLaMA base models - continual pretraining LLaMA2

This phase aims to adapt LLaMA2 models to better understand and generate text relevant to the medical context using the pre-training datasets we constructed. The training involves sequences of medical texts, where the model learned to predict the next token in a sequence, maximizing the likelihood, where is the parameter set of LLaMA2 models. This training was executed on the University of Florida’s HiPerGator AI supercomputer with 160 A100 80GB GPUs. We employed the AdamW optimizer with hyperparameters set to 0.9 and to 0.95, alongside a weight decay of 0.00001 and a learning rate of  $8e-6$ . We used a cosine learning rate scheduler with a 0.05 warmup ratio for gradual adaptation to training complexity and bf16 precision for computational efficiency. Gradient accumulation was set to 16 steps, and training was limited to one epoch. We utilized DeepSpeed<sup>26</sup> for model parallelism.

**Me-LLaMA chat models - instruction fine-tuning Me-LLaMA:** We further fine-tuned Me-LLaMA base models, using the developed 214k instruction samples. The training objective is to maximize the likelihood:

$\mathcal{L}(\Theta) = \operatorname{argmax} \sum_{(x^i, y^i) \in (X, Y)} \log p(y^i | x^i; \Theta)$ , where  $x^i$  represents the input instruction,  $y^i$  is the ground truth response, and  $\Theta$  is the parameter set of Me-LLaMA. Executed using 8 A100 GPUs, the fine-tuning process was set to run for 3 epochs with a learning rate of 1e-5. We used a weight decay of 0.00001 and a warmup ratio of 0.01 for regularization and gradual learning rate increase. We utilized LoRA-based<sup>27</sup> parameter-efficient fine-tuning.

## Evaluation Benchmark

**Biomedical and clinical NLP tasks:** Existing studies<sup>2,3,10,11</sup> in the medical domain have primarily focused on evaluating the QA task. In this study, we build an extensive medical evaluation benchmark (MIBE), encompassing six critical text analysis tasks: QA, NER, RE, Text Classification, Text Summarization and NLI. These tasks collectively involve 12 datasets meticulously sourced from biomedical, and clinical domains as shown in Table 3.

Table 3  
Details of data splits and evaluation metrics of each dataset in the evaluation benchmark.

Data	Task	Train	Valid	Test	Evaluation
PubMedQA* <sup>28</sup>	QA	190,143	21,126	500	Accuracy, Macro-F1
MedQA <sup>29</sup>	QA	10,178	1,272	1,273	Accuracy, Macro-F1
MedMCQA* <sup>30</sup>	QA	164,540	18,282	4,183	Accuracy, Macro-F1
EmrQA <sup>31</sup>	QA	122,326	30,581	26,804	Exact match, F1
i2b2 <sup>32</sup>	NER	6,0875	7,400	7,451	Entity-level Macro-F1
DDI <sup>33</sup>	RE	18,779	7,244	5,761	Macro-F1
HoC <sup>34</sup>	Classification	1,108	157	315	Label-wise Macro-F1
MTSample <sup>35</sup>	Classification	4,999	500	999	Accuracy, Macro-F1
PubMed <sup>36</sup>	Summarization	117,108	6,631	6,658	Rouge, BERTScore
MIMIC-CXR <sup>17</sup>	Summarization	122,014	957	1,606	Rouge, BERTScore
BioNLI <sup>37</sup>	NLI	5,544	5,000	6,308	Accuracy, Macro-F1
MedNLI <sup>38</sup>	NLI	11,232	1,422	1,395	Accuracy, Macro-F1

### Complex clinical case diagnosis task

we further assessed the effectiveness of Me-LLaMA in diagnosing complex clinical cases, a critical task given the increasing burden of diseases and the need for timely and accurate diagnosis to support clinicians. Recent studies demonstrate that LLMs have the potential to address this challenge.<sup>39</sup> Specifically, we evaluated the diagnostic accuracy of Me-LLaMA on 70 challenging medical cases from the New England Journal of Medicine clinicopathologic conferences (NEJM CPCs) published between January 2021 and December 2022, as collected from an existing study.<sup>39</sup> The NEJM CPCs are well-known for their unique and intricate clinical cases, which have long been used as benchmarks for evaluating challenging medical scenarios. In line with previous research,<sup>39,40</sup> we employed automatic evaluations based on top-K (where k = 1,2,3,4,5) accuracy, defined as the percentage of cases where the correct diagnosis appeared within the top-K positions of the differential diagnosis list predicted by the assessed models. We utilized GPT-4o, a state-of-the-art (SOTA) LLM, to automatically assess whether each diagnosis from the model's differential diagnosis list matched the gold standard final diagnosis, consistent with these prior studies. Existing studies<sup>40</sup> have shown that LLM-based automatic calculation of top-K accuracy is comparable to human evaluation. Besides automatic evaluation, we had a clinician specializing in internal medicine perform a manual evaluation of top-k accuracy (k = 1, 5). For more details on data processing, automatic evaluation, and human evaluation, see Appendix A.3.

## Evaluation Settings

We evaluated Me-LLaMA at two evaluation settings including zero-shot and supervised learning to evaluate their performance and generalization ability across various tasks compared to baseline models.

## Supervised Learning

In the supervised learning setting, we evaluated Me-LLaMA 13/70B base models' performances adapted to downstream tasks. We conducted the task-specific finetuning on Me-LLaMA base models (Me-LLaMA task-specific) with each training set of assessed datasets in Table 6, and then assessed the performance of Me-LLaMA task-specific models on test datasets. We employed the AdamW optimizer. For datasets with fewer than 10,000 training samples, we fine-tuned the models for 5 epochs, while for larger datasets, the fine-tuning was conducted for 3 epochs. A uniform learning rate of  $1e-5$  was used across all datasets. Our baseline models including LLaMA2 Models (7B/13B/70B)<sup>7</sup>: they are open-sourced LLMs released by Meta AI. PMC-LLaMA 13B<sup>2</sup> is a biomedical LLM continually pre-trained on biomedical papers and medical books. Meditron7B/70B<sup>10</sup>: they are medical LLMs based on LLaMA2-7B/70B, continual pre-trained with a mix of clinical guidelines, medical papers and abstracts.

## Zero-shot Learning

We assessed our Me-LLaMA 13/70B-chat models' zero-shot learning capabilities, which are key for new task understanding and response without specific prior training. We compared our models and baseline models' zero-shot, using standardized prompts (detailed in Table A.2 shown in Appendix 0.2) for each test dataset from Table 2. We compared Me-LLaMA 13/70B-chat models with the following baseline models: ChatGPT/GPT-4<sup>4,5</sup>: SOTA commercialized LLMs. We used the version of "gpt-3.5-turbo-0301" for ChatGPT, and the version of "gpt-4-0314" for GPT-4. LLaMA2-7B/13B/70B-chat<sup>7</sup> models were adaptations of the LLaMA2 series, optimized for dialogue and conversational scenarios. Medalpaca-7B/13B<sup>3</sup> models were based on LLaMA-7B/13B, specifically fine-tuned for tasks in the medical domain. The PMC-LLaMA-13B-chat<sup>2</sup> model is an instruction-tuned medical LLM based on PMC-LLaMA-13B. The AlpaCare-13B<sup>13</sup> model is specifically tailored for clinical tasks based on LLaMA-2 13B by instruction tuning. Meditron 70B<sup>10</sup> is a medical LLM, continually pre-trained with a mix of clinical guidelines, biomedical papers, and abstracts based on LLaMA2 70B.

## RESULTS

### Overall Performance: Medical Text Analysis

Table 4 compares the performance of our Me-LLaMA 13/70B foundation models against other open LLMs in the supervised setting. We can observe that the Me-LLaMA 13B model surpassed the similar-sized medical foundation model PMC-LLaMA 13B on 11 out of 12 datasets and outperformed the general foundation model LLaMA2 13B on 10 out of 12 datasets. Moreover, it is noticed that the Me-LLaMA 13B model was competitive with LLaMA2 70B and Meditron 70B, which have significantly larger parameter sizes, on 8 out of 12 datasets. As for 70B models, Me-LLaMA 70B achieved the best performance on 9 out of 12 datasets, when benchmarked against LLaMA2 70B and Meditron 70B.



Table 4  
The supervised fine-tuning performance of various open source LLMs on six tasks.

Task	Dataset	Metric	LLaMA2 13B	PMC-LLaMA 13B	Me-LLaMA 13B	LLaMA2 70B	Meditron 70B	Me-LLaMA 70B
Question answering	PubMedQA	Acc	0.800	0.778	<b>0.802</b>	0.800	0.800*	<b>0.814</b>
		Macro-F1	0.560	0.544	<b>0.562</b>	0.560	-	<b>0.572</b>
	MedQA	Acc	0.467	0.456	<b>0.493</b>	0.598	0.607*	<b>0.623</b>
		Macro-F1	0.465	0.454	<b>0.487</b>	0.595	-	<b>0.621</b>
	MedMCQA	Acc	0.527	0.548	<b>0.557</b>	0.626	<b>0.651*</b>	0.643
		Macro-F1	0.524	0.545	<b>0.551</b>	0.625	-	<b>0.640</b>
	EmrQA	Acc	0.789	0.810	<b>0.857</b>	0.847	0.850	<b>0.854</b>
		F1	0.730	0.738	<b>0.751</b>	0.751	0.751	<b>0.751</b>
Named entity recognition	i2b2	Macro-F1	0.904	0.901	<b>0.906</b>	<b>0.913</b>	0.908	0.910
Relation extraction	DDI	Macro-F1	<b>0.622</b>	0.622	0.559	0.746	0.737	<b>0.779</b>
Classification	HoC	Macro-F1	<b>0.696</b>	0.422	0.684	0.818	0.702	<b>0.841</b>
	MTsample	Macro-F1	0.430	0.345	<b>0.451</b>	0.458	0.284	<b>0.544</b>
Summarization	PubMed	R-L	0.191	0.091	<b>0.197</b>	<b>0.211</b>	0.197	0.209
		BERTS	0.663	0.516	<b>0.679</b>	0.689	0.677	<b>0.700</b>
	MIMIC-CXR	R-L	0.437	0.139	<b>0.453</b>	0.440	0.458	<b>0.476</b>
		BERTS	0.816	0.694	<b>0.821</b>	0.813	0.824	<b>0.828</b>
Natural language inference	BioNLI	Macro-F1	0.409	0.332	<b>0.447</b>	0.447	0.444	<b>0.566</b>
	MedNLI	Macro-F1	0.881	0.868	<b>0.903</b>	0.884	0.897	<b>0.916</b>

\*The performance of Meditron 70B on the PubMedQA, MedQA, and MedMCQA datasets is cited from the meditron paper<sup>10</sup> to have a fair comparison.

Table 5 shows the zero-shot performance of Me-LLaMA chat models and other instruction tuned open LLMs with chat ability on various tasks. Among 13B models, Me-LLaMA 13B-chat outperformed LLaMA2 13B-chat, PMC-LLaMA-chat, Medalpaca 13B in almost all 12 datasets. Me-LLaMA outperformed AlpaCare-13B in 9 out of 12 datasets. Among models with 70B parameters, Me-LLaMA 70B-chat consistently outperformed LLaMA2-70B-chat on 11 out of 12 datasets. It is worth noting that Me-LLaMA13B-chat showed better performance than LLaMA2-70B-chat—a model with a significantly larger parameter size—on 6 out of 12 datasets and was competitive with the LLaMA2-70B-chat in 3 out of 6 remaining datasets.

Table 5  
The zero-shot performance of various open source LLMs with chat capability.

Task	Dataset	Metric	LLaMA2-13B-chat	PMC-LLaMA-chat	Medalpaca-13B	AlpaCare-13B	Me-LLaMA 13B-chat	LLaMA2-70B-chat	Me-LLaMA 70B-chat
Question answering	PubMedQA	Accuracy	0.546	0.504	0.238	0.538	<b>0.700</b>	0.668	<b>0.768</b>
		Macro-F1	0.457	0.305	0.192	0.373	<b>0.504</b>	0.477	<b>0.557</b>
	MedQA	Accuracy	0.097	0.207	0.143	0.304	<b>0.427</b>	0.376	<b>0.523</b>
		Macro-F1	0.148	0.158	0.102	0.281	<b>0.422</b>	0.367	<b>0.521</b>
	MedMCQA	Accuracy	0.321	0.212	0.205	0.385	<b>0.449</b>	0.339	<b>0.539</b>
		Macro-F1	0.243	0.216	0.164	0.358	<b>0.440</b>	0.273	<b>0.538</b>
	EmrQA	Accuracy	0.001	<b>0.053</b>	0.000	0.001	0.048	0.050	<b>0.119</b>
		F1	0.098	0.304	0.040	0.198	<b>0.307</b>	0.251	<b>0.346</b>
Named entity recognition	i2b2	Macro-F1	0.143	0.091	0.000	<b>0.173</b>	0.166	0.321	<b>0.329</b>
Relation extraction	DDI	Macro-F1	0.090	0.147	0.058	0.110	<b>0.214</b>	0.087	<b>0.283</b>
Classification	HoC	Macro-F1	0.228	0.184	0.246	0.267	<b>0.335</b>	0.309	<b>0.544</b>
	MTsample	Macro-F1	0.133	0.083	0.003	<b>0.273</b>	0.229	0.254	<b>0.384</b>
Summarization	PubMed	Rouge-L	0.161	0.028	0.014	<b>0.167</b>	0.116	<b>0.192</b>	0.169
		BERTS*	0.671	0.128	0.117	<b>0.671</b>	0.445	<b>0.684</b>	0.678
	MIMIC-CXR	Rouge-L	0.144	0.139	0.010	0.134	<b>0.400</b>	0.131	<b>0.418</b>
		BERTS*	0.704	0.694	0.502	0.702	<b>0.797</b>	0.696	<b>0.787</b>
Natural language inference	BioNLI	Macro-F1	0.173	0.159	0.164	0.170	<b>0.195</b>	0.297	<b>0.436</b>
	MedNLI	Macro-F1	0.412	0.175	0.175	0.275	<b>0.472</b>	0.515	<b>0.675</b>

\*BERTS: BERTScore.<sup>41</sup>

Figure 2 further compares the performance of Me-LLaMA models in the zero-shot and supervised learning setting, against ChatGPT and GPT-4. Due to privacy concerns, which preclude the transmission of clinical datasets with patient information to ChatGPT and GPT-4, we conducted our comparison across 8 datasets that are not subject to these limitations. The results of ChatGPT and GPT-4 on three QA datasets are referenced from the OpenAI's paper.<sup>1</sup> We compared the Rouge-1<sup>42</sup> score for the summarization dataset PubMed, the accuracy score for three QA datasets, and the Macro-F1 score for the remaining datasets. With task-specific supervised fine-tuning, Me-LLaMA models surpassed ChatGPT on 7 out of 8 datasets and excelled GPT-4 on 5 out of 8 datasets. In the zero-shot setting, Me-LLaMA models outperformed ChatGPT on 5 datasets; but it fell short on 7 datasets, when compared with GPT-4. It's crucial to highlight that Me-LLaMA's model size is significantly smaller—13/70B parameters versus at least 175B for ChatGPT and GPT-4. Despite this size discrepancy, Me-LLaMA models have showcased an impressive performance and a strong ability for supervised learning and zero-shot learning across a broad spectrum of medical tasks, underscoring its efficiency and potential in the field.

## Clinical Application: Complex Clinical Case Diagnosis

Figure 3 shows the top-K ( $1 \leq K \leq 5$ ) accuracy of Me-LLaMA-70B-chat, ChatGPT, GPT-4, and LLaMA2-70B-chat, in the complex clinical case diagnosis task. We can see Me-LLaMA-70B-chat model achieved comparable performance with GPT-4 and ChatGPT, and significantly outperforms LLaMA2-70B-chat. The human evaluation result in Fig. 4 again shows that Me-LLaMA-70B-chat outperformed GPT-4 in both top-1 and top-5 accuracy. These results demonstrated the potential of Me-LLaMA models for challenging clinical applications.

## **Ablation Study: Impact of Continual Pretraining and Instruction Tuning**

Table 6 compares the zero-shot performances of Me-LLaMA models and their backbone models LLaMA2, to illustrate the impact of continual pre-training and instruction tuning. Table 3 clearly demonstrates that both continual pre-training and instruction tuning significantly enhanced the zero-shot capabilities of models. For example, the Me-LLaMA 70B model showed an improvement in performance ranging from 2.1–55% across various datasets in comparison to the LLaMA2 13B model, highlighting the benefits of continual pre-training. The instruction tuning was also found to provide great increases in zero-shot performance. For instance, the Me-LLaMA-70B-chat model displayed enhancements in performance from 3.7–41.9% relative to the Me-LLaMA 70B foundation model, which had not undergone instruction tuning. This enhancement suggests the critical role of instruction finetuning in boosting the model's ability to leverage context in learning tasks, even without supervised fine-tuning and prior examples.

Table 6  
The comparison of zero-shot performances among Me-LLaMA models and their backbone models LLaMA2.

Dataset	Metric	LLaMA2 13B (backbone)	Me-LLaMA 13B  (backbone + pre-train)	Me-LLaMA-13B- chat (backbone + pre-train + instruction tuning)	LLaMA2 70B  (backbone)	Me-LLaMA 70B  (backbone + pre-train)	Me-LLaMA- 70B-chat  (backbone + pre-train + instruction tuning)
PubMedQA	Acc	0.216	0.266	<b>0.700</b>	0.132	0.682	<b>0.768</b>
	Macro-F1	0.177	0.250	<b>0.504</b>	0.152	0.520	<b>0.557</b>
MedQA	Acc	0.000	0.000	<b>0.427</b>	0.005	0.281	<b>0.523</b>
	Macro-F1	0.000	0.000	<b>0.422</b>	0.009	0.350	<b>0.521</b>
MedMCQA	Acc	0.003	0.003	<b>0.449</b>	0.012	0.447	<b>0.539</b>
	Macro-F1	0.006	0.005	<b>0.440</b>	0.024	0.396	<b>0.538</b>
EmrQA	Acc	0.000	0.005	<b>0.048</b>	0.000	0.021	<b>0.119</b>
	F1	0.038	0.122	<b>0.307</b>	0.000	0.172	<b>0.346</b>
i2b2	Macro-F1	0.008	0.030	<b>0.263</b>	0.181	0.224	<b>0.329</b>
DDI	Macro-F1	0.035	0.036	<b>0.214</b>	0.034	0.118	<b>0.283</b>
HoC	Macro-F1	0.253	0.210	<b>0.335</b>	0.255	0.252	<b>0.544</b>
MTsample	Macro-F1	0.042	0.072	<b>0.229</b>	0.066	0.226	<b>0.384</b>
PubMed	R-L	<b>0.170</b>	0.168	0.116	0.167	0.119	<b>0.169</b>
	BERTS	<b>0.654</b>	0.654	0.445	0.654	0.654	<b>0.678</b>
MIMIC-CXR	R-L	0.051	0.172	<b>0.400</b>	0.059	0.137	<b>0.418</b>
	BERTS	0.566	0.697	<b>0.797</b>	0.577	0.649	<b>0.787</b>
BioNLI	Macro-F1	0.109	0.060	<b>0.195</b>	0.285	<b>0.499</b>	0.436
MedNLI	Macro-F1	0.172	0.206	<b>0.472</b>	0.265	0.256	<b>0.675</b>

## DISCUSSION

### Model Performance

We introduced a novel medical LLM family including, Me-LLaMA 13B and Me-LLaMA 70B, which encode comprehensive medical knowledge, along with their chat-optimized variants: Me-LLaMA-13/70B-chat, with strong zero-shot learning ability, for medical applications. These models were developed through the continual pre-training and instruction tuning of LLaMA2 models, using the largest and most comprehensive biomedical and clinical data. Compared to existing studies, we perform the most comprehensive evaluation, covering six critical text analysis tasks. Our evaluations reveal that Me-LLaMA models outperform existing open-source medical LLMs in various learning scenarios, showing less susceptibility to catastrophic forgetting and achieving competitive results against major commercial models including ChatGPT and GPT-4. Our work paves the way for more accurate, reliable, and comprehensive medical LLMs, and underscores the potential of LLMs on medical applications.

In the zero-shot setting, medical LLMs including GPT-4 displayed low performance on certain tasks, e.g., NER and RE, which are also noted by other studies.<sup>43,44</sup> When compared with other NLP tasks with higher performance, we noticed that one of the main reasons for low performance is that LLMs' responses often lacked the conciseness and precision expected, with instances of missing outputs noted. The unexpected outputs also cause significant challenges to automatic evaluation metrics. Therefore, more investigation is needed to further improve medical LLMs' performance across tasks in the zero-shot setting<sup>31</sup> and enhance the automatic assessment of these medical LLMs' zero-shot capabilities. For the complex clinical case diagnosis, the Me-LLaMA-chat model had competitive performance and even outperformed GPT-4 in human evaluation. Existing studies have demonstrated GPT-4 is arguably one of the strongest LLMs in this task.<sup>45</sup> The robust performance of Me-LLaMA showed potential in assisting challenging clinical applications. It is noticed that variations in test sizes and evaluation methods across different studies contribute to the observed differences in performance between GPT-4 in our paper and other studies. We also noted that both the Me-LLaMA-chat model and GPT-4 faced difficulties identifying the correct diagnosis within the top ranks, underscoring the difficulty of this task. Additionally, while the NEJM CPCs offer a rigorous test for these models, they do not encompass the full range of a physician's duties or broader clinical competence. Therefore, complex clinical diagnosis remains a challenging area that demands more effective models and improved evaluation benchmarks to better capture the complexities of real-world clinical scenarios.

## **Model Development**

During our model development, we noticed the importance of diversity of the data sources during the pre-training and instruction-tuning phases. Our empirical results revealed that the PMC-LLaMA 13B model, which employed a data mix ratio of 19:1 between medical and general domain data, exhibited around 2.7% performance drop across both general and biomedical tasks. On the other hand, the Meditron models, 7B, and 70B, with a 99:1 mix ratio, demonstrated improvements in biomedical tasks, yet they still saw around 1% declines in the performance of general tasks. In contrast, our models, which adopt a 4:1 ratio, have shown enhancements in their performance for both general and medical tasks. This suggests that the integration of general domain data plays a vital role in mitigating the knowledge-forgetting issue during pre-training.<sup>11,24,25</sup> However, determining the optimal balance between general domain data and specialized medical data is nontrivial, requiring careful empirical analysis. Future studies should examine methods to better determine the optimal ratio.

Our model development also underscores the balance between cost and effectiveness in pre-training versus instruction tuning of LLMs. Pre-training, exemplified by the LLaMA2 70B model, is notably resource-heavy, requiring about 700 hours on 160 A100 GPUs per epoch. Conversely, instruction tuning is far less resource-demanding, needing roughly 70 hours on 8 A100 GPUs per epoch, making it much more affordable than pre-training. Despite this, instruction tuning alone enhanced the performance of the Me-LLaMA-13B-chat model, achieving improvements ranging from 12% to 45% across 11 out of 12 datasets when compared to its backbone model – Me-LLaMA 13B, in the zero-shot setting. This efficiency advocates for prioritizing instruction tuning in scenarios with limited resources, highlighting its potential for cost-effective model enhancement.

## **Use of Me-LLaMA Models**

The Me-LLaMA models, available in both 13B and 70B sizes, as well as in base and chat-optimized versions, unlock a wide array of medical applications, guided by the crucial balance between model size and resource availability. The base models serve as robust foundations with extensive medical knowledge, adaptable through supervised fine-tuning for specialized tasks. Conversely, the chat versions excel in instruction-following ability and zero-shot learning, making them highly effective in zero-shot or few-shot learning scenarios. Larger models, like the 70B, provide deeper understanding and more complex reasoning abilities, ideal for comprehensive medical analyses. Yet, their deployment requires significant computing resources, posing challenges in resource-limited settings. On the other hand, the 13B models offer a practical compromise, balancing efficiency with effectiveness, thus ensuring broader accessibility for various applications. Our findings indicate that the Me-LLaMA 13B achieves performance on par with the 70B variant across most datasets, suggesting its viability for diverse medical tasks where computational or financial resources are a concern.

## **Limitations**

It is crucial to acknowledge the limitations of the current versions of Me-LLaMA models. Like all existing LLMs, they are susceptible to generating information with factual errors or biased information. To mitigate this, future studies could incorporate methodologies

like reinforcement learning from human feedback (RLHF).<sup>46</sup> This approach could align the models' responses more closely with human values and ensure they are grounded in factual medical knowledge. Another limitation is the current token handling capacity, capped at 4096 tokens, which is a constraint inherited from the backbone LLaMA2 model. Addressing this limitation could involve extending the models' capability to handle longer contexts. This could be achieved by integrating advanced attention techniques, such as sparse local attention,<sup>47</sup> that are able to handle extensive contexts.

## Declarations

### DATA AVAILABILITY

All datasets employed in the continual pre-training process and evaluation are accessible from their original published venues. The PubMed Central and PubMed Abstracts subset from The Pile are available at <https://huggingface.co/datasets/EleutherAI/pile>. MIMIC-IV and MIMIC-CXR datasets can be accessed under the PhysioNet Credentialed Health Data Use Agreement 1.5.0 at <https://physionet.org/content/mimic-iv-note/2.2/> and <https://physionet.org/content/mimic-cxr/2.0.0/> respectively. The RedPajama data is open-released at <https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1>. Alpaca data is openly released at: [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca). Dolly data is openly released at: <https://huggingface.co/datasets/databricks/databricks-dolly-15k>. Share GPT data can be accessed at: [https://huggingface.co/datasets/anon8231489123/ShareGPT\\_Vicuna\\_unfiltered](https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered). The clinical instruction tuning data based on MIMIC-IV and MIMIC-CXR can be accessed under the PhysioNet Credentialed Health Data Use Agreement 1.5.0 through: <https://huggingface.co/clinicalnlp>. The Medical Flash Cards and wikidoc QA datasets can be accessed at <https://huggingface.co/medalpaca>. Other remaining instruction tuning data can be openly accessed at: <https://huggingface.co/clinicalnlp>. Me-LLaMA 13B and Me-LLaMA 70B models can be accessed at: <https://physionet.org/content/me-llama/1.0.0/>, subject to the completion of a credentialed health data use agreement.

### CODE AVAILABILITY

The code used for evaluation is available at: <https://github.com/BIDS-Xu-Lab/Me-LLaMA>.

### ACKNOWLEDGEMENT

This work received support from the National Institutes of Health (NIH) under grant numbers: 1RF1AG072799, 1R01AG078154, R01AG073435, R01LM013519, RF1AG084178, R01AG083039, R01CA284646, R01AI172875, R01AG080991, R01AG080624, R01AG080429, 1K99LM01402, 1K99LM014614-01, NIH/NCATS UL1 TR001427, CDC U18 DP006512, and Patient-Centered Outcomes Research Institute (PCORI) under grant numbers: PCORI RI-FLORIDA-01-PS1, PCORI ME-2018C3-14754. We express our sincere appreciation to the creators of datasets such as the MIMIC, the Pile, and RedPajama for making these valuable resources available to the research community. We extend our gratitude to the UF Research Computing team, under the leadership of Dr. Erik Deumens, for their generous provision of computational resources through the UF HiperGator-AI cluster.

### AUTHOR CONTRIBUTION

QX contributed to the conceptualization of the study, conducted the literature search, developed the methodology, contributed to the software development, carried out validation processes, and was primarily responsible for writing the original draft of the manuscript. QC contributed to the conceptualization of the study, developed the methodology, and contributed to reviewing and editing the manuscript. AC played a key role in data curation and project administration, overseeing the planning and execution of research activities, and contributing to reviewing and editing the manuscript. CP, YH, FL, XP, JH, JZ, VK, XZ and LQ were instrumental in software development and validation and reviewing the manuscript. HH took charge of visualization, specifically in the preparation of figures to support the study's findings, involved in the discussion and reviewing the manuscript. LOM and YW were involved in the discussion, review, and editing of the paper. DS was involved in the discussion, human evaluation, and reviewing the manuscript. HX and JB provided overall supervision for the project, including study design, execution, and evaluation, coordination of study team and resources, and thorough review and revision of the manuscript. All authors reviewed the manuscript critically for scientific content, and all authors gave final approval of the manuscript for publication.

### COMPETING INTEREST

The authors have no financial or non-financial conflicts of interest to disclose.

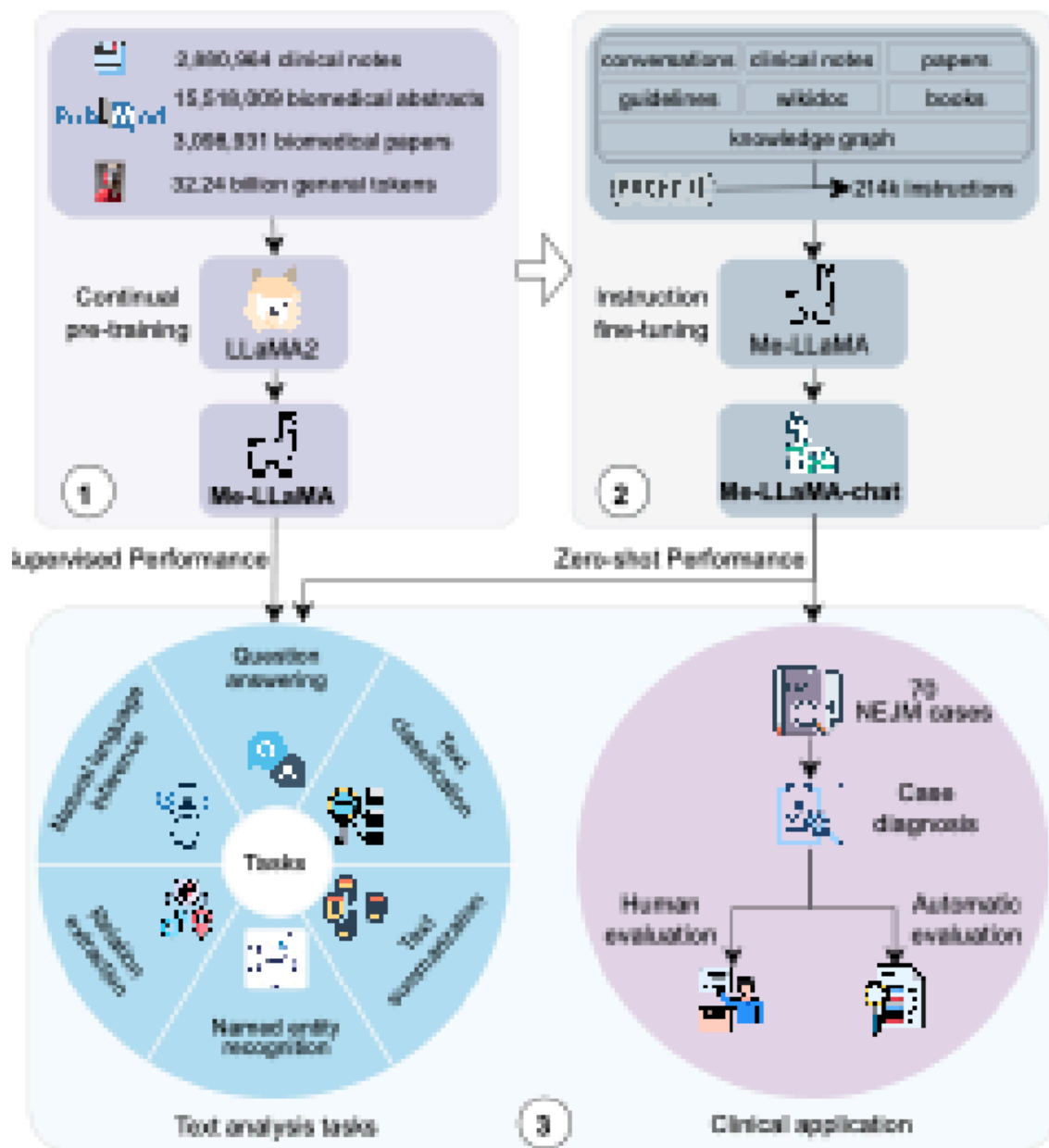
## References

1. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023).
2. Wu, C., Zhang, X., Zhang, Y., Wang, Y. & Xie, W. Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454 (2023).
3. Han, T. et al. Medalpaca—an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247 (2023).
4. <https://openai.com/blog/chatgpt>
5. OpenAI. Gpt-4 technical report (2023). 2303.08774.
6. Touvron, H. et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).
7. Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).
8. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023).
9. Peng, C. et al. A study of generative large language model for medical research and healthcare. arXiv preprint arXiv:2305.13523 (2023).
10. Chen, Z. et al. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079 (2023).
11. Gema, Aryo, et al. Parameter-efficient fine-tuning of llama for the clinical domain. arXiv preprint arXiv:2307.03042 (2023).
12. Li, Y. et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* 15 (2023).
13. Zhang, X. et al. Alpacare: Instruction-tuned large language models for medical application. arXiv preprint arXiv:2310.14558 (2023).
14. Gao, L. et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027 (2020).
15. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1–9 (2016).
16. Johnson, A. et al. MIMIC-IV. *PhysioNet*. Available online at: <https://physionet.org/content/mimiciv/1.0/> (accessed August 23, 2021) (2020).
17. Johnson, A. E. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data* 6, 317 (2019).
18. Computer, T. Redpajama: an open dataset for training large language models (2023).
19. Taori, R. et al. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca) (2023).
20. Conover, M. et al. Free dolly: Introducing the world’s first truly open instruction-tuned llm (2023).
21. Zheng, L. et al. Judging llm-as-a-judge with mt-bench and chatbot arena (2023). 2306.05685.
22. Ben Abacha, A., Shivade, C. & Demner-Fushman, D. Overview of the MedQA 2019 shared task on textual inference, question entailment and question answering. In *ACL-BioNLP 2019* (2019).
23. Ben Abacha, A. et al. Bridging the gap between consumers’ medication questions and trusted answers. In *MEDINFO 2019* (2019).
24. Abacha, A. B., Agichtein, E., Pinter, Y. & Demner-Fushman, D. Overview of the medical question answering task at TREC 2017 liveqa. In *TREC*, 1–12 (2017).
25. Yu, B., Li, Y. & Wang, J. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4664–4674 (2019).
26. <https://github.com/microsoft/DeepSpeed>
27. Hu, E. J. et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (2022).

28. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. Pubmedqa: A dataset for biomedical research question answering. arXiv preprint arXiv:1909.06146 (2019).
29. Zhang, X., Wu, J., He, Z., Liu, X. & Su, Y. Medical exam question answering with large-scale reading comprehension. In Proceedings of the AAAI conference on artificial intelligence, vol. 32 (2018).
30. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Conference on Health, Inference, and Learning, 248–260 (PMLR, 2022).
31. Pampari, A., Raghavan, P., Liang, J. & Peng, J. emrqa: A large corpus for question answering on electronic medical records. arXiv preprint arXiv:1809.00732 (2018).
32. Uzuner, Özlem, et al. "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text." Journal of the American Medical Informatics Association 18.5 (2011): 552-556.
33. Segura-Bedmar, I., Martínez Fernández, P. & Herrero Zazo, M. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013) (Association for Computational Linguistics, 2013).
34. Baker, S. et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics 32, 432–440 (2016).
35. <https://www.kaggle.com/code/ritheshsreenivasan/clinical-text-classification>
36. Cohan, A. et al. A discourse-aware attention model for abstractive summarization of long documents. arXiv preprint arXiv:1804.05685 (2018).
37. Bastan, M., Surdeanu, M. & Balasubramanian, N. Bionli: Generating a biomedical nli dataset using lexico-semantic constraints for adversarial examples. arXiv preprint arXiv:2210.14814 (2022).
38. Romanov, A. & Shivade, C. Lessons from natural language inference in the clinical domain. arXiv preprint arXiv:1808.06752 (2018).
39. Kanjee, Zahir, Byron Crowe, and Adam Rodman. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. Jama 330.1 (2023): 78-80.
40. McDuff, Daniel, et al. Towards accurate differential diagnosis with large language models. arXiv preprint arXiv:2312.00164 (2023).
41. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019).
42. Lin, CY. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, 74–81 (2004).
43. Chen, Qingyu, et al. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:2305.16326 (2023).
44. Hu, Yan, et al. Improving large language models for clinical named entity recognition via prompt engineering. Journal of the American Medical Informatics Association (2024): ocad259.
45. Savage, Thomas, et al. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digital Medicine 7.1 (2024): 20.
46. Stiennon, N. et al. Learning to summarize with human feedback. Advances Neural Information Processing Systems 33, 3008–3021 (2020).
47. Chen, Y. et al. Longlora: Efficient fine-tuning of long-context large language models. arXiv preprint arXiv:2309.12307 (2023).

## Figures





**Figure 1**

Overview of the study. Our study has three main components including pre-training, instruction fine-tuning and evaluation. Pre-training: we firstly developed the Me-LLaMA base models by continual pre-training LLaMA2 with 129 billion tokens from mixed pre-training text data. Instruction fine-tuning: Me-LLaMA-chat models were further developed by instruction tuning Me-LLaMA base models with 214K instructions. Evaluation: Finally, we evaluated the Me-LLaMA base models in a supervised learning setting across six text analysis tasks, and the Me-LLaMA-chat models in a zero-shot setting on both text analysis tasks and a clinical diagnosis task.

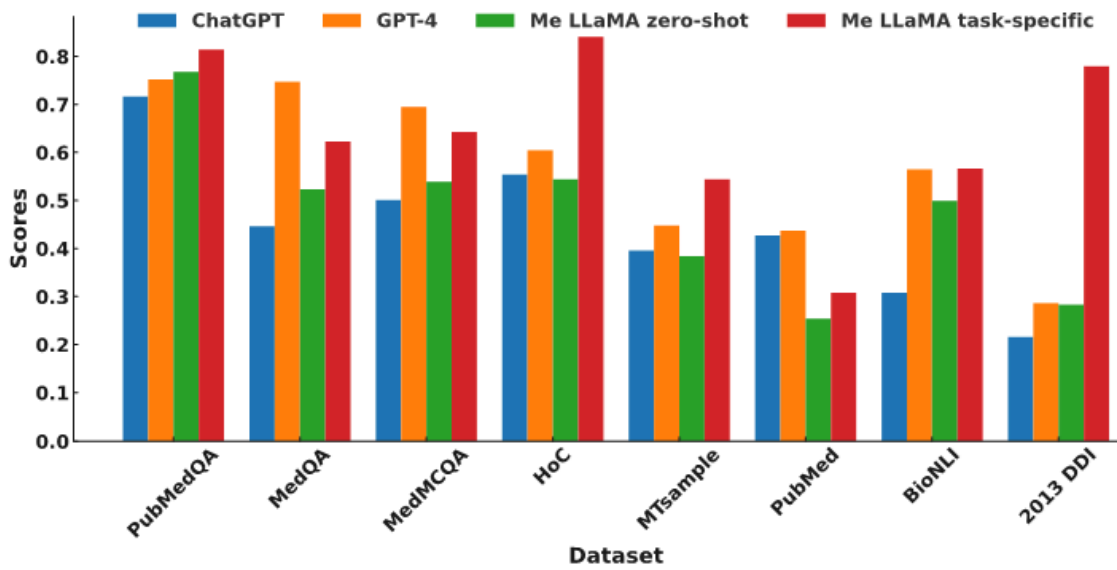


Figure 2

The performance comparison of Me-LLaMA models in both zero-shot (Me-LLaMA zero-shot) and supervised learning (Me-LLaMA task-specific) settings, against the zero-shot performance of ChatGPT and GPT-4.

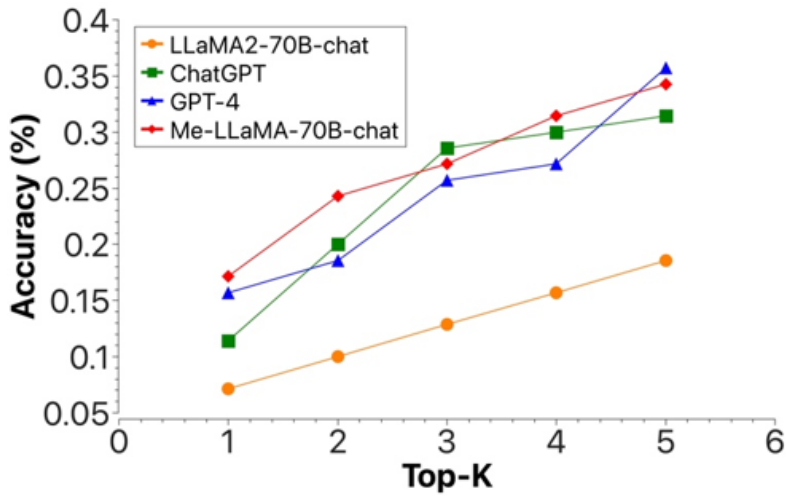
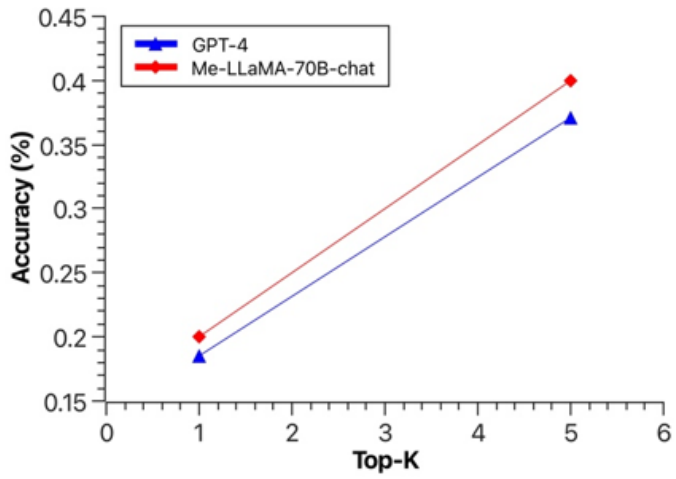


Figure 3

The top-k (1 ≤ k ≤ 5) accuracy of different LLMs in complex clinical case diagnosis, with automatic evaluation.



**Figure 4**

The top-1 and top-5 accuracy of Me-LLaMA-70B-chat and GPT-4 in complex clinical case diagnosis, with human evaluation.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [APPENDIX.docx](#)