npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01533-1

Medical foundation large language models for comprehensive text analysis and beyond

Check for updates

Qianqian Xie[1,5], Qingyu Chen[1,5], Aokun Chen[2,5], Cheng Peng[2], Yan Hu[3], Fongci Lin[1], Xueqing Peng[1], Jimin Huang[1], Jeffrey Zhang[1], Vipina Keloth[1], Xinyu Zhou[1], Lingfei Qian[1], Huan He[1], Dennis Shung[1,4], Lucila Ohno-Machado[1], Yonghui Wu[2], Hua Xu[1] ✉ & Jiang Bian[2] ✉

Recent advancements in large language models (LLMs) show significant potential in medical applications but are hindered by limited specialized medical knowledge. We present Me-LLaMA, a family of open-source medical LLMs integrating extensive domain-specific knowledge with robust instruction-following capabilities. Me-LLaMA is developed through continual pretraining and instruction tuning of LLaMA2 models using diverse biomedical and clinical data sources (e.g., biomedical literature and clinical notes). We evaluated Me-LLaMA on six text analysis tasks using 12 benchmarks (e.g., PubMedQA and MIMIC-CXR) and assessed its clinical utility in complex case diagnosis through automatic and human evaluations. Me-LLaMA outperforms existing open medical LLMs in zero-shot and supervised settings and surpasses ChatGPT and GPT-4 after task-specific instruction tuning for most text analysis tasks. Its performance is also comparable to ChatGPT and GPT-4 for diagnosing complex clinical cases. Our findings highlight the importance of combining domain-specific continual pretraining with instruction tuning to enhance performance in medical LLMs.

Large language models (LLMs) have shown great potential in improving medical applications such as clinical documentation, diagnostic accuracy, and patient care management[1–3]. However, general-domain LLMs often lack specialized medical knowledge because they are primarily trained on non-medical datasets[4], limiting their effectiveness in healthcare settings. Although commercial LLMs, such as ChatGPT and GPT-4[4], 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[5,6]; Yet these models still fall short due to their general-domain training[7–9].

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[3], ChatDoctor[10], and AlpaCare[11], 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[8], 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[2], Meditron[9], and Clinical LLaMA[12].

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[2] 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

[1]Department of Biomedical Informatics and Data Science, Yale School of Medicine, Yale University, New Haven, CT, USA. [2]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA. [3]School of Biomedical Informatics, University of Texas Health Science, Center at Houston, Houston, TX, USA. [4]Department of Medicine (Digestive Diseases), Yale School of Medicine, Yale University, New Haven, CT, USA. [5]These authors contributed equally: Qianqian Xie, Qingyu Chen, Aokun Chen. ✉e-mail: hua.xu@yale.edu; bianji@iu.edu

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). (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[6] models. Leveraging an extensive medical dataset—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. Prior studies[2,3,7–12] have primarily focused on evaluating the QA task. For example, PMC-LLaMA[2] and Meditron[9] evaluated their model performance on medical QA tasks derived from domain-specific literature, while MedAlpaca[3] and ChatDoctor[10] focused on conversational QA. In contrast, we conduct a comprehensive evaluation 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.

Results

Overall performance of medical text analysis

Table 1 compares the performance of our Me-LLaMA 13/70B foundation models against other open LLMs in the supervised setting. The performance of Meditron 70B on the PubMedQA, MedQA, and MedMCQA datasets is cited from the meditron paper to have a fair comparison. 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 2 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.

Figure 1 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

Table 1 | 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	0.802	0.800	0.800	0.814
		Macro-F1	0.560	0.544	0.562	0.560	–	0.572
	MedQA	Acc	0.467	0.456	0.493	0.598	0.607	0.623
		Macro-F1	0.465	0.454	0.487	0.595	–	0.621
	MedMCQA	Acc	0.527	0.548	0.557	0.626	0.651	0.643
		Macro-F1	0.524	0.545	0.551	0.625	–	0.640
	EmrQA	Acc	0.789	0.810	0.857	0.847	0.850	0.854
		F1	0.730	0.738	0.751	0.751	0.751	0.751
Named entity recognition	i2b2	Macro-F1	0.904	0.901	0.906	0.913	0.908	0.910
Relation extraction	DDI	Macro-F1	0.622	0.622	0.559	0.746	0.737	0.779
Classification	HoC	Macro-F1	0.696	0.422	0.684	0.818	0.702	0.841
	MTsample	Macro-F1	0.430	0.345	0.451	0.458	0.284	0.544
Summarization	PubMed	R-L	0.191	0.091	0.197	0.211	0.197	0.209
		BERTS	0.663	0.516	0.679	0.689	0.677	0.700
	MIMIC-CXR	R-L	0.437	0.139	0.453	0.440	0.458	0.476
		BERTS	0.816	0.694	0.821	0.813	0.824	0.828
Natural language inference	BioNLI	Macro-F1	0.409	0.332	0.447	0.447	0.444	0.566
	MedNLI	Macro-F1	0.881	0.868	0.903	0.884	0.897	0.916

BERTS means BERTScore[28].

Table 2 | 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	0.700	0.668	0.768
		Macro-F1	0.457	0.305	0.192	0.373	0.504	0.477	0.557
	MedQA	Accuracy	0.097	0.207	0.143	0.304	0.427	0.376	0.523
		Macro-F1	0.148	0.158	0.102	0.281	0.422	0.367	0.521
	MedMCQA	Accuracy	0.321	0.212	0.205	0.385	0.449	0.339	0.539
		Macro-F1	0.243	0.216	0.164	0.358	0.440	0.273	0.538
	EmrQA	Accuracy	0.001	0.053	0.000	0.001	0.048	0.050	0.119
		F1	0.098	0.304	0.040	0.198	0.307	0.251	0.346
Named entity recognition	i2b2	Macro-F1	0.143	0.091	0.000	0.173	0.166	0.321	0.329
Relation extraction	DDI	Macro-F1	0.090	0.147	0.058	0.110	0.214	0.087	0.283
Classification	HoC	Macro-F1	0.228	0.184	0.246	0.267	0.335	0.309	0.544
	MTsample	Macro-F1	0.133	0.083	0.003	0.273	0.229	0.254	0.384
Summarization	PubMed	Rouge-L	0.161	0.028	0.014	0.167	0.116	0.192	0.169
		BERTS	0.671	0.128	0.117	0.671	0.445	0.684	0.678
	MIMIC-CXR	Rouge-L	0.144	0.139	0.010	0.134	0.400	0.131	0.418
		BERTS	0.704	0.694	0.502	0.702	0.797	0.696	0.787
Natural language inference	BioNLI	Macro-F1	0.173	0.159	0.164	0.170	0.195	0.297	0.436
	MedNLI	Macro-F1	0.412	0.175	0.175	0.275	0.472	0.515	0.675

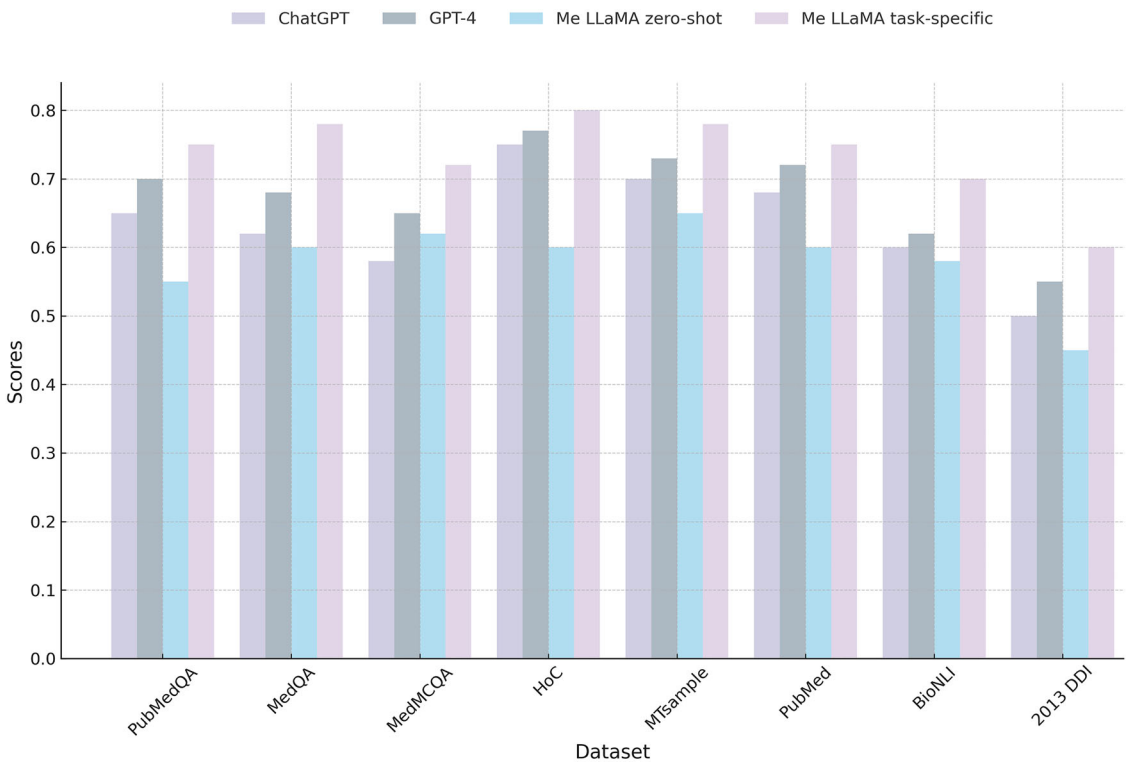

Fig. 1 | Performance comparison of Me-LLaMA models with ChatGPT and GPT-4. The figure presents the zero-shot performance of Me-LLaMA (Me-LLaMA zero-shot) alongside its supervised learning performance (Me-LLaMA task-specific), compared against the zero-shot performance of ChatGPT and GPT-4 across 8 datasets.

paper[1]. We compared the Rouge-1[13] 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

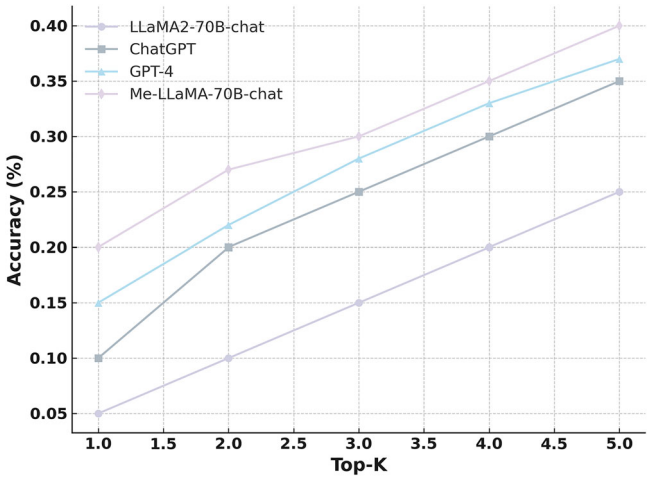

Fig. 2 | Model performance in the complex clinical case diagnosis task under automatic evaluation. The figure presents the top-K accuracy (where $1 \le K \le 5$) of Me-LLaMA-70B-chat, ChatGPT, GPT-4, and LLaMA2-70B-chat on a complex clinical case diagnosis task, evaluated automatically.

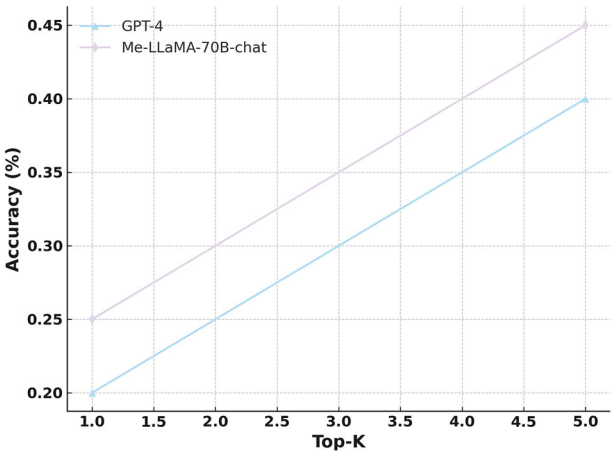

Fig. 3 | Model performance in the complex clinical case diagnosis task under human evaluation. The figure shows the top-1 and top-5 accuracy of Me-LLaMA/-70B-chat and GPT-4 in a complex clinical case diagnosis task, evaluated through human assessment.

and zero-shot learning across a broad spectrum of medical tasks, underscoring its efficiency and potential in the field.

Performance of complex clinical case diagnosis

Figure 2 shows the top-K ($1 \le K \le 5$) 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. 3 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.

Impact of continual pretraining and instruction tuning

Table 3 demonstrates the impact of continual pre-training and instruction tuning on zero-shot performance across medical NLP tasks. It clearly demonstrates that both continual pre-training and instruction tuning significantly enhanced the zero-shot capabilities of models. Instruction tuning alone provides significant performance improvements over the base LLaMA2 models, as seen in LLaMA2 13B, where accuracy on PubMedQA increases from 0.216 to 0.436. This suggests that instruction tuning is highly effective in enhancing the model's ability to follow task-specific prompts. In contrast, continual pre-training on medical data yields relatively modest improvements, particularly for smaller models. Me-LLaMA 13B shows only slight gains over LLaMA2 13B, likely due to the smaller scale of domain-specific pre-training data compared to LLaMA2's original training corpus, which exceeds 2 T tokens. Additionally, continual pre-training may not provide as strong of a task-specific signal as instruction tuning, limiting its impact in zero-shot settings. However, for larger models like Me-LLaMA 70B, continual pre-training results in more notable improvements, with performance gains ranging from 2.1% to 55% across various datasets, demonstrating its value in capturing specialized domain knowledge. The best results are consistently achieved when both continual pre-training and instruction tuning are applied together, as seen in Me-LLaMA-70B-chat, which outperforms all other configurations. This indicates that while instruction tuning is the most efficient approach for improving task performance, continual pre-training provides a complementary boost, particularly for larger models where additional domain adaptation enhances overall effectiveness.

Discussion

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.

Despite these strengths, we observed certain challenges in specific tasks, such as NER and RE[14,15], where even advanced models like GPT-4 exhibited low performance. 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 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[16]. 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.

Our results also emphasize the importance of data diversity during model development. 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,

Table 3 | 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 only)	LLaMA2 13B-instruct (backbone + instruction tuning only)	Me-LLaMA-13B-chat (backbone + pre-train + instruction tuning)	LLaMA2 70B (backbone)	Me-LLaMA 70B (backbone + pre-train only)	LLaMA2 70B-instruct (backbone + instruction tuning only)	Me-LLaMA-70B-chat (backbone + pre-train + instruction tuning)
PubMedQA	Acc	0.216	0.266	0.436	0.700	0.132	0.682	0.764	0.768
	Macro-F1	0.177	0.250	0.416	0.504	0.152	0.520	0.531	0.557
MedQA	Acc	0.000	0.000	0.013	0.427	0.005	0.281	0.499	0.523
	Macro-F1	0.000	0.000	0.024	0.422	0.009	0.350	0.493	0.521
MedMCQA	Acc	0.003	0.003	0.014	0.449	0.012	0.447	0.501	0.539
	Macro-F1	0.006	0.005	0.029	0.440	0.024	0.396	0.493	0.538
EmrQA	Acc	0.000	0.005	0.050	0.048	0.000	0.021	0.181	0.119
	F1	0.038	0.122	0.286	0.307	0.000	0.172	0.399	0.346
i2b2	Macro-F1	0.008	0.030	0.232	0.263	0.181	0.224	0.245	0.329
DDI	Macro-F1	0.035	0.036	0.164	0.214	0.034	0.118	0.121	0.283
HoC	Macro-F1	0.253	0.210	0.194	0.335	0.255	0.252	0.563	0.544
MTsample	Macro-F1	0.042	0.072	0.176	0.229	0.066	0.226	0.364	0.384
PubMed	R-L	0.170	0.168	0.183	0.116	0.167	0.119	0.112	0.169
	BERTS	0.654	0.654	0.667	0.445	0.654	0.654	0.601	0.678
MIMIC-CXR	R-L	0.051	0.172	0.360	0.400	0.059	0.137	0.367	0.418
	BERTS	0.566	0.697	0.791	0.797	0.577	0.649	0.784	0.787
BioNLI	Macro-F1	0.109	0.060	0.185	0.195	0.285	0.499	0.345	0.436
MedNLI	Macro-F1	0.172	0.206	0.457	0.472	0.265	0.256	0.657	0.675

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[2,9]. 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.

The cost-effectiveness of instruction tuning is another important consideration. Pre-training, exemplified by the LLaMA2 70B model, is notably resource-heavy, requiring about 160*700 GPU hours per epoch. Conversely, instruction tuning is far less resource-demanding, needing roughly 8*70 GPU hours per epoch, making it much more affordable than pre-training. While continual pre-training aims to incorporate specialized medical knowledge into the model, the observed performance improvements, particularly for smaller models like Me-LLaMA 13B, are relatively modest. This limited improvement can be attributed to several factors. First, the amount of domain-specific pre-training data used is significantly smaller compared to the original pre-training data of LLaMA2, which exceeds 2 T tokens. This discrepancy suggests that larger amounts of domain-specific data may be required to fully activate the model's potential. Second, the continual pre-training strategy itself could be optimized further. As noted earlier, the process faces challenges such as catastrophic forgetting, where the model loses general-domain knowledge during adaptation to specialized data. Despite this, models trained only with the instruction tuning demonstrate that instruction tuning alone can significantly enhance performance at a fraction of the computational cost. This highlights instruction tuning as a practical and cost-effective alternative, particularly in scenarios where computational resources are limited.

The Me-LLaMA models, available in both 13B and 70B sizes, as well as in base and chat-optimized versions, enable a wide array of medical applications, guided by the crucial balance between model size and resource availability. The base models provide a strong foundation for supervised fine-tuning on specialized tasks, while the chat-optimized versions excel in instruction-following and zero-shot scenarios. Larger models, like the 70B, deliver superior reasoning capabilities but require significant computational resources, making the 13B models a practical alternative for broader accessibility. Notably, the Me-LLaMA 13B model achieves performance comparable to its 70B counterpart across most datasets, demonstrating its utility for diverse medical tasks in resource-limited settings. These features suggest that Me-LLaMA models could be explored for various medical applications. Potential areas of use include: (1) clinical decision support, where these models might assist in analyzing patient records, generating differential diagnoses, and synthesizing medical literature to support evidence-based decision-making; (2) medical education, where chat-optimized versions could serve as interactive tools for teaching medical students and trainees by providing explanations for complex medical topics and simulating diagnostic reasoning; and (3) administrative tasks, where these models may help streamline workflows by summarizing clinical notes and generating discharge summaries, potentially reducing the documentation burden on clinicians. Further research and evaluation are warranted to assess Me-LLaMA's real-world effectiveness and limitations in these clinical application settings.

Despite these advancements, it is crucial to acknowledge the current Me-LLaMA models still have certain limitations that require further attention. 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)[17]. 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[18], that are able to handle extensive contexts.

Additionally, while MIMIC is the largest publicly available EHR dataset, its size is still relatively small compared to other data sources, which

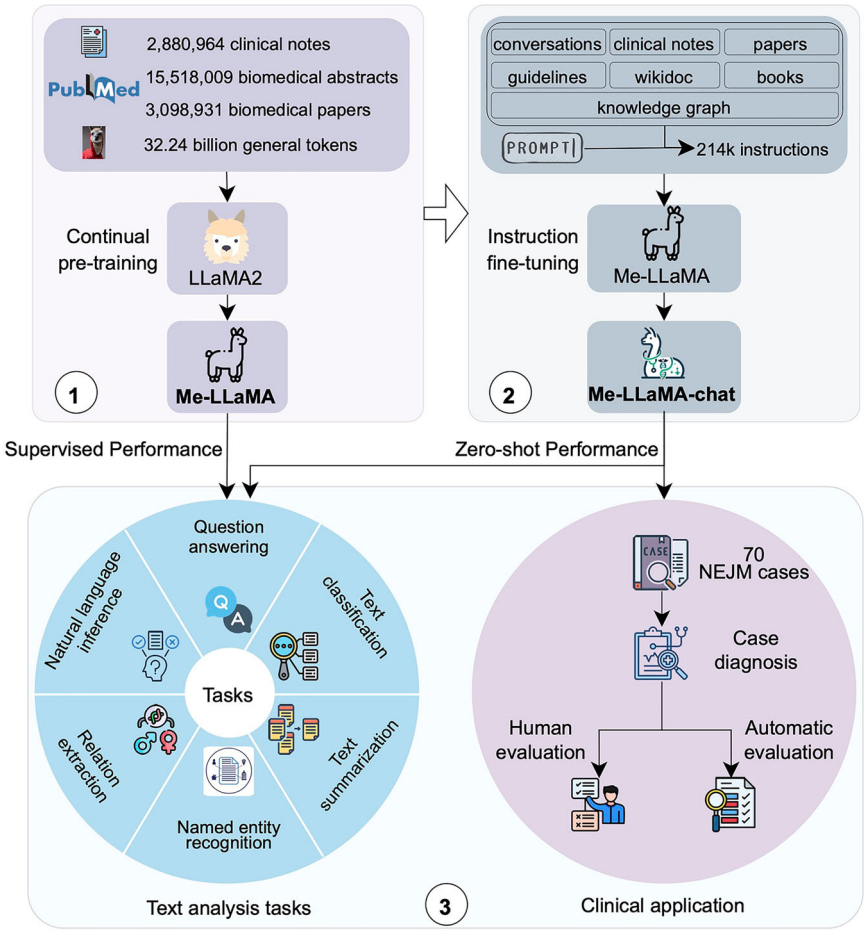

Fig. 4 | Overview of the study. Our study has three main components including pre-training, instruction fine-tuning and evaluation. Pre-training: we first developed the Me-LLaMA base models by continual pre-training LLaMA2 with 129 billion tokens from mixed pre-training text data. Instruction fine-tuning: Me-LLaMA-chat models were further developed by instruction-tuning Me-LLaMA base models with 214 K 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.

Table 4 | 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[3]	LLaMA	7/13B	✓	✗	-	160 K	QA	04/14/2023
ChatDoctor[12]	LLaMA2	7B	✓	✗	-	100 K	QA	05/24/2023
AlpaCare[28]	LLaMA	7/13B	✓	✗	-	52 K	QA, Summarization	10/23/2023
Clinical LLaMA[11]	LLaMA	7B	✗	✓	-	-	Classification	07/06/2023
Meditron[10]	LLaMA2	7/70B	✓	✗	48B	-	QA	11/27/2023
PMC-LLaMA[2]	LLaMA	7/13B	✓	✗	79B	514 K	QA	04/27/2023
Me-LLaMA	LLaMA2	13/70B	✓	✓	129B	214 K	QA, NER, RE, Classification, Summarization, NLI, Medical Diagnosis	06/05/2024

may impact Me LLaMA's generalizability on real-world scenarios. This limitation stems primarily from privacy concerns surrounding clinical data, which significantly restrict the availability of large-scale EHR datasets. In this study, we included only MIMIC in the training of Me LLaMA because it is readily available for public dissemination through the established MIMIC data access procedures. Moving forward, we plan to train Me LLaMA on much larger proprietary clinical datasets, such as EHRs from Yale New Haven Health and the University of Florida Health. However, the terms of distribution and dissemination for models trained on such proprietary data will need to be carefully negotiated with our institutions' data governance committees to ensure the safety and confidentiality of clinical data.

Methods

We utilized LLaMA2[6] as the backbone model and developed Me-LLaMA through the process of continual pre-training and instruction tuning of LLaMA2, using 129B tokens and 214 K instruction tuning samples from general, biomedical, and clinical domains. Figure 4 shows an overview of our study. Table 4 presents the comparison of Me-LLaMA models and existing open source medical LLMs.

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. Our dataset integrates a vast collection of biomedical literature from PubMed Central and PubMed Abstracts, sourced from the Pile dataset[19]. The PubMed Central subset includes 3,098,931 biomedical articles, and the PubMed Abstracts section encompasses abstracts from 15,518,009 documents. This comprehensive biomedical dataset provides a rich source of medical knowledge and research findings. To incorporate

Table 5 | The overall instruction tuning dataset

Task	Type	Source	Size	Copy right
General	Conversation	Alpaca[29]	20,000	CC-BY-NC 4.0
		Dolly[30]		CC-BY-SA-3.0
		ShareGPT[31]		Apache-2.0
Biomedical	Conversation	HealthCareMagic[10]	20,000	Reserved by HealthCareMagic and Icliniq
		Icliniq[10]		
	Instructions	MedInstruct[11]	52,000	CC BY-NC 4.0
	Question Answering	Medical Flash Cards[3]	34,000	No commercialized use
		MEDIQA[32]	2,220	CC BY 4.0
		MedicationQA[33]	690	CC BY 4.0
		LiveQA[34]	634	CC BY 4.0
		WikiDocPatient[3]	5490	CC BY-SA 4.0
		GuidelineQA	2000	Common Crawl (other)
	Summarization	PubMed Central	10,000	CC BY
	Next Sentence Generation	PubMed Central	20,000	CC BY
	Key words prediction	PubMed Central	10,000	CC BY
	Causal Relation Detection	PubMed[35]	2450	CC BY
	Relation Extraction	UMLS knowledge graph[2]	10,000	Openrail
Clinical	QA, summarization, classification, mortality prediction	MIMIC-III[20], MIMIC-IV[21]	30,000	PhysioNet credentialed health data use agreement 1.5.0

real-world clinical scenarios and reasoning, we included de-identified free-text clinical notes from MIMIC-III[20], MIMIC-IV[21], and MIMIC-CXR[22]. MIMIC-III contains 112,000 clinical reports records. MIMIC-IV contains 331,794 de-identified discharge summaries and 2,321,355 radiology reports. MIMIC-CXR adds further depth with 227,835 radiology reports for radiographic studies. Moreover, to prevent the model from forgetting acquired general knowledge, we incorporated a subset from the RedPajama[23] dataset, a replication of LLaMA2's pre-training data. This dataset is composed of diverse data slices, including processed CommonCrawl dumps, GitHub data, scientific articles from arXiv, a subset of Wikipedia pages, and popular websites from StackExchange. Our 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 5. 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 the Supplementary Information, Supplementary Table 1.

Training details

As shown in Fig. 3, we developed the Me-LLaMA 13B and 70B base models by continually 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.

The first phase aims to develop Me-LLaMA base models, and adapt LLaMA2 models to better understand and generate text relevant to the medical context using the pre-training datasets we constructed. The objective is to enhance the model's ability to understand and generate domain-specific text by optimizing it to predict the next word in a sequence based on the preceding context. 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 β_1 to 0.9 and β_2 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[24] for model parallelism.

We further fine-tuned Me-LLaMA base models to develop Me-LLaMA chat models, using the developed 214k instruction samples. In this phase, the models are trained to produce accurate and contextually appropriate responses to specific input instructions. 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[25] parameter-efficient fine-tuning.

Evaluation benchmark

Existing studies[2,3,9] 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 6.

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[26]. 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[27]. 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[26,27], 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

Table 6 | Details of data splits and evaluation metrics of each dataset in the evaluation benchmark

Data	Task	Train	Valid	Test	Evaluation
PubMedQA[36]	QA	190,143	21,126	500	Accuracy, Macro-F1
MedQA[37]	QA	10,178	1272	1273	Accuracy, Macro-F1
MedMCQA[38]	QA	164,540	18,282	4183	Accuracy, Macro-F1
EmrQA[39]	QA	122,326	30,581	26,804	Exact match, F1
i2b2[40]	NER	6,0875	7400	7451	Entity-level Macro-F1
DDI[41]	RE	18,779	7244	5761	Macro-F1
HoC[42]	Classification	1108	157	315	Label-wise Macro-F1
MTSample (https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions)	Classification	4999	500	999	Accuracy, Macro-F1
PubMed[43]	Summarization	117,108	6631	6658	Rouge, BERTScore
MIMIC-CXR[22]	Summarization	122,014	957	1606	Rouge, BERTScore
BioNLI[44]	NLI	5544	5000	6308	Accuracy, Macro-F1
MedNLI[45]	NLI	11,232	1422	1395	Accuracy, Macro-F1

whether each diagnosis from the model's differential diagnosis list matched the gold standard final diagnosis, consistent with these prior studies. Existing studies[27] 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 the Supplementary Information.

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)[6]: they are open-sourced LLMs released by Meta AI. PMC-LLaMA 13B[2] is a biomedical LLM continually pre-trained on biomedical papers and medical books. Meditron7B/70B[9]: these are medical LLMs based on LLaMA2-7B/70B, continually 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 the Supplementary Information, Supplementary Table 2) for each test dataset from Table 2. We compared Me-LLaMA 13/70B-chat models with the following baseline models: ChatGPT/GPT-4[4]: 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[6] models were adaptations of the LLaMA2 series, optimized for dialogue and conversational scenarios. Medalpaca-7B/13B[3] models were based on LLaMA-7B/13B, specifically fine-tuned for tasks in the medical domain. The PMC-LLaMA-13B-chat[2] model is an instruction-tuned medical LLM based on PMC-LLaMA-13B. The AlpaCare-13B[11] model is specifically tailored for clinical tasks based on LLaMA-2 13B by instruction tuning. Meditron 70B[9] is a medical LLM, continually pre-trained with a mix of clinical guidelines, biomedical papers, and abstracts based on LLaMA2 70B.

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. 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. 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/clinicalnlplab. 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/clinicalnlplab. 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.

Received: 14 November 2024; Accepted: 20 February 2025;
Published online: 05 March 2025

References

1. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Preprint at https://doi.org/10.48550/arXiv.2303.13375 (2023).
2. Wu, C. et al. PMC-LLaMA: toward building open-source language models for medicine. *J. Am. Med Inf. Assoc.* **31**, 1833–1843 (2024).
3. Han, T. et al. MedAlpaca - An Open-Source Collection of Medical Conversational AI Models and Training Data. Preprint at https://doi.org/10.48550/arXiv.2304.08247 (2023).
4. Achiam, O. J. et al. GPT-4 Technical Report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2023).
5. Touvron, H. et al. LLaMA: Open and Efficient Foundation Language Models. Preprint at https://doi.org/10.48550/arXiv.2302.13971 (2023).

6. Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at https://doi.org/10.48550/arXiv.2307.09288 (2023).
7. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2022).
8. Peng, C. A. I. et al. A study of generative large language model for medical research and healthcare. *NPJ Digital Medicine*. **6** (2023).
9. Chen, Z. et al. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2311.16079 (2023).
10. 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).
11. Zhang, X. et al. AlpaCare: Instruction-tuned Large Language Models for Medical Application. Preprint at https://doi.org/10.48550/arXiv.2310.14558 (2023).
12. Gema, A., Minervini, P., Daines, L., Hope, T. & Alex, B. Parameter-Efficient Fine-Tuning of LLaMA for the Clinical Domain. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*. 91-104. Mexico City, Mexico (Association for Computational Linguistics, 2024).
13. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. 74-81. Barcelona, Spain (Annual Meeting of the Association for Computational Linguistics, 2004).
14. Chen, Q. et al. A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations. Preprint at https://doi.org/10.48550/arXiv.2305.16326 (2023).
15. Hu, Y. et al. Improving large language models for clinical named entity recognition via prompt engineering. *J. Am. Med. Inform. Assoc. JAMIA* **31**, 1812–1820 (2023).
16. Savage, T., Nayak, A., Gallo, R., Rangan, E. S. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digit. Med*. **7** (2023).
17. Stiennon, N. et al. Learning to summarize with human feedback. *Advances in neural information processing systems*. (2020).
18. Chen, Y. et al. LongLoRA: Efficient Fine-tuning of Long-Context Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2309.12307 (2023).
19. Gao, L. et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. Preprint at https://doi.org/10.48550/arXiv.2101.00027 (2020).
20. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data*. **3** (2016).
21. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data*. **10** (2023).
22. Johnson, A. E. W. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data*. **6** (2019).
23. Weber, M. et al. RedPajama: an Open Dataset for Training Large Language Models. *Advances in neural information processing systems*. (2025).
24. Rasley, J., Rajbhandari, S., Ruwase, O. & He, Y. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. (2020).
25. Hu, J. E. et al. LoRA: Low-Rank Adaptation of Large Language Models. In *The Twelfth International Conference on Learning Representations.* (2022).
26. Kanjee, Z., Crowe, B. & Rodman, A. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA*. (2023).
27. McDuff, D. et al. Towards Accurate Differential Diagnosis with Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2312.00164 (2023).
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. (2019).
29. Taori, R. et al. Stanford Alpaca: An Instruction-following LLaMA model https://github.com/tatsu-lab/stanford_alpaca (2023).
30. Conover, M. et al. Free dolly: Introducing the world's first truly open instruction-tuned llm https://huggingface.co/datasets/databricks/databricks-dolly-15k (2023).
31. Zheng, L. et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*. (2023).
32. Abacha, A. B., Shivade, C. P. & Demner-Fushman, D. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. 370-379. Florence, Italy (Association for Computational Linguistics, 2019).
33. Abacha, A. B. et al. Bridging the gap between consumers' medication questions and trusted answers. *Stud. health Technol. Inform.* **264**, 25–29 (2019).
34. Abacha, A. B., Agichtein, E., Pinter, Y. & Demner-Fushman, D. Overview of the Medical Question Answering Task at TREC 2017 LiveQA. In (Text Retrieval Conference, 2017).
35. 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. Hong Kong, China (Association for Computational Linguistics, 2019).
36. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W. & Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. 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)*. 2567-2577. Hong Kong, China (Association for Computational Linguistics, 2019).
37. 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*. (2018).
38. Pal, A., Umapathi, L. K. & Sankarasubbu, M. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*. 248-260. (Proceedings of Machine Learning Research, 2022).
39. Pampari, A., Raghavan, P., Liang, J. J. & Peng, J. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2357-2368. Brussels, Belgium (2018).
40. Uzuner, Ö., South, B. R., Shen, S. & Duvall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* **18 5**, 552–556 (2011).
41. Segura-Bedmar, I., Martínez, P. & Herrero-Zazo, M. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 341-350. Atlanta, Georgia, USA (Association for Computational Linguistics, 2013).
42. Baker, S. et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics* **32**, 432–440 (2016).
43. Cohan, A. et al. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 615-621. New Orleans, Louisiana (Association for Computational Linguistics, 2018).
44. Bastan, M., Surdeanu, M. & Balasubramanian, N. BioNLI: Generating a Biomedical NLI Dataset Using Lexico-semantic Constraints for

Adversarial Examples. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. 5093-5104. Abu Dhabi, United Arab Emirates (Association for Computational Linguistics, 2022).
45. Romanov, A. & Shivade, C. P. Lessons from Natural Language Inference in the Clinical Domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 1586-1596. Brussels, Belgium (Association for Computational Linguistics, 2018).

Acknowledgements

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 contributions

Q.X. 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. Q.C. contributed to the conceptualization of the study, developed the methodology, and contributed to reviewing and editing the manuscript. A.C. 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. C.P., Y.H., F.L., X.P., J.H., J.Z., V.K., X.Z., and L.Q. were instrumental in software development and validation and reviewing the manuscript. H.H. took charge of visualization, specifically in the preparation of figures to support the study's findings, involved in the discussion and reviewing the manuscript. D.S. was involved in the discussion, human evaluation, and reviewing the manuscript. L.O.M. and Y.W. were involved in the discussion, review, and editing of the paper. H.X and J.B. 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 interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01533-1.

Correspondence and requests for materials should be addressed to Hua Xu or Jiang Bian.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025