Lite-Me-LLaMA: Resource-Efficient Large Language Models for Medical Applications

Qianqian Xie, PhD[1], Aokun Chen, PhD[2], Cheng Peng, PhD[2], Lingfei Qian, PhD[1], Yan Wang, PhD[1], Xuguang Ai, MS[1], Jimin Huang, PhD[1], Rui Shi[1], Gui Yang[1], Dennis Shung, PhD[3], Qingyu Chen, PhD[1], Yonghui Wu, PhD[1], Jiang Bian, PhD[1], and Hua Xu, PhD[1]

[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]Department of Medicine (Digestive Diseases), Yale School of Medicine, Yale University, New Haven, CT, USA

ABSTRACT

Objective: To overcome the challenges of high costs, infrastructure demands, and data privacy concerns in medical applications of large language models (LLMs), we introduce Lite-Me-LLaMA, a suite of lightweight, open-source medical LLMs. By enhancing LLaMA3-8B with extensive medical pre-training and instruction fine-tuning, it is designed to be efficient, transparent, and readily deployable in clinical settings, requiring minimal computational resources for seamless integration into workflows while maintaining robust data privacy and security.

Materials and Methods: We developed Lite-Me-LLaMA by continually pre-training the LLaMA3-8B model on a 72.47 billion token corpus sourced from biomedical literature, clinical guidelines, and textbooks. This equips the base model with comprehensive medical knowledge and enables it to be efficiently adapted to specific medical tasks through supervised fine-tuning using just one NVIDIA A100 40GB GPU. We then created Lite-Me-LLaMA-Instruct by fine-tuning the base model with 602,000 instruction-response pairs from diverse biomedical sources, optimized for robust performance on unseen medical tasks and instructions. It can be deployed for real-time applications through prompt engineering without additional training, using a single 16GB consumer-grade GPU.

Results: Lite-Me-LLaMA achieved state-of-the-art performance among lightweight open-source LLMs (≤8B parameters) and larger LLMs (13B parameters) across three commonly used medical question-answering benchmarks, in both supervised learning and zero-shot learning settings.

Discussion: The superior performance proves that Lite-Me-LLaMA models offer a resource-efficient, locally deployable tool that ensures data privacy and transparency. It allows fine-tuning with minimal resources, enabling clinicians to tailor the model to specific tasks like diagnostics. Additionally, Lite-Me-LLaMA-Instruct can be deployed instantly without retraining, streamlining workflows and enhancing decision-making efficiency while maintaining low costs and privacy

Conclusion: We introduce Lite-Me-LLaMA as a valuable contribution to medical AI, offering a resource-efficient, open-source tool designed for real-world clinical settings. This work demonstrates that combining continual pre-training with instruction fine-tuning creates high-performing models that are both efficient and effective. Lite-Me-LLaMA empowers healthcare institutions with limited computational resources to leverage cutting-edge technologies for improved clinical workflows and decision-making. We openly release our models and evaluation code, and are committed to ongoing development based on the most advanced backbone LLMs, such as LLaMA 3.2.

Key words: Medical Large Language Models, Lightweight Large Language Models, Continual Pre-training, Instruction Fine-tuning

1 INTRODUCTION

There is a pressing need for solutions that provide clinicians with timely and accurate medical information, enabling evidence-based decision-making in fast-paced clinical environments.[1–3] Clinicians frequently face complex decisions under pressure, often with limited or conflicting information and insufficient time,[4] encountering an average of two clinical questions for every three patients seen.[5–7] However, current AI-driven clinical question-answering (QA) systems fall short in language expressivity and interactive capabilities,[8] leaving a significant

gap between their abilities and the needs of clinicians. Despite advancements in deep learning that have improved system performance,[9–11] these systems remain constrained by narrow generalization abilities, failing to fully streamline clinical workflows and enhance patient outcomes as required in real-world medical practice.[12]

Recently, large language models (LLMs) have been introduced, offering a transformative approach to AI-based clinical decision support and showing significant potential to revolutionize medical practice.[13] Advanced close-source models, such as ChatGPT[1] and GPT-4[14] trained on general domain broad topics, can handle complex medical questions and generate human-like responses. These closed-source models, with proprietary architecture and data, use in-context learning to perform new tasks without specific training, sometimes even outperforming physicians in delivering personalized advice.[15–17] In contrast, open-source models like Meditron[18] and Me-LLaMA,[19] enriched with domain-specific clinical and biomedical data, offer greater flexibility and consistently outperform general-domain models like GPT-4 in specialized medical tasks.

Although LLMs show significant potential in healthcare, their integration in real-world workflows with clinicians faces challenges related to high costs, infrastructure demands, and data privacy concerns.[20] Closed-source models like GPT-4 and MedGemini[21] offer quick, ready-to-use solutions but come with steep API fees ranging from $0.03 to $0.06 per 1,000 tokens, potentially amounting to $2700 monthly.[22] These models cannot be fine-tuned locally, limiting customization, and their reliance on external servers raises concerns about data privacy and HIPAA compliance.[23,24] Open-source models like Meditron and Me-LLaMA offer more flexibility, supporting local fine-tuning and deployment to mitigate privacy risks.[25,26] However, they require substantial resources, i.e, fine-tuning large models with up to 70 billion parameters using NVIDIA A100 80GB GPUs can cost $2,304 for 192 hours with a large scale dataset,[27,28] along with significant infrastructure and technical expertise for hosting.

To address it, we introduce *Lite-Me-LLaMA*, a suite of lightweight, open-source medical LLMs designed for local deployment with minimal GPU memory requirements. We begin by continual pre-training the LLaMA3-8B model on 72.47 billion tokens sourced from biomedical literature, clinical guidelines, and books, creating the Lite-Me-LLaMA base model, which bridges the medical knowledge gap. It can be adapted to specific medical tasks by supervised fine-tuning, which requires only one NVIDIA A100 40GB GPU, making it feasible for institutions with modest infrastructure. Building on this foundation, we developed Lite-Me-LLaMA-Instruct, fine-tuning the base model with 602,000 instruction-response pairs from diverse biomedical sources, enhancing the model's ability to handle new medical tasks without requiring additional training. It is designed to be deployed with one 16GB consumer-grade GPU. This allows for real-time inference in resource-constrained environments while maintaining strong performance on diverse medical tasks. Our models have demonstrated superior performance over other lightweight open-source medical LLMs (smaller than 8B parameters) and even outperform larger models (13B parameters) on three commonly used medical question-answering benchmarks, while maintaining low resource demands. Our findings also highlight the value of continual pre-training and instruction fine-tuning in improving LLM performance on medical tasks, making Lite-Me-LLaMA an ideal choice for healthcare settings that require both efficiency and privacy. By making the Lite-Me-LLaMA models publicly available[2] and releasing our evaluation code[3], we aim to empower the healthcare and research communities to advance accessible, secure AI solutions. We are committed to continually maintaining and developing advanced, lightweight open-source models that prioritize accessibility, privacy, and performance for real-world clinical applications.

2 METHOD

We developed Lite-Me-LLaMA by continual pre-training and instruction fine-tuning of the LLaMA3-8B model.

2.1 Lite-Me-LLaMA: Enhance LLaMA3 with Medical Knowledge by Continual Pre-training

2.1.1 Continual Pre-training Corpus

To effectively adapt the general-purpose LLaMA3-8B model to the medical domain, we developed a mixed continual pre-training dataset that equips the model with comprehensive medical knowledge and enhances its performance on medical tasks. Our continual pre-training corpus was meticulously curated to ensure comprehensive coverage of medical knowledge, drawing from a wide array of high-quality medical data sources (as shown in Table 1): (1) Biomedical Literature: The corpus includes the full texts of 3,098,931 biomedical articles from the PubMed Central (PMC) subset, covering publications from the year 2000 to June 2020. In addition, abstracts from 15,518,009 biomedical articles in the PubMed online repository were also incorporated. Both of these resources were sourced from The Pile dataset.[29] (2) NIH Reporter: We integrated 939,668 NIH grant abstracts from awarded applications spanning the fiscal years 1985 to 2020.[29] These abstracts, sourced from

[1] `https://openai.com/index/chatgpt/`

[2] `https://huggingface.co/YBXL`

[3] `https://github.com/BIDS-Xu-Lab/Lite-Me-LLaMA`

the NIH Reporter, provide valuable insights into cutting-edge biomedical research topics. (3) Clinical Guidelines: Our corpus also includes 37,000 clinical practice guidelines collected from nine authoritative online medical sources.[18] These guidelines offer critical, evidence-based recommendations for clinical care, helping to bolster the model's knowledge of best practices in various medical specialties and real-world clinical decision-making. (4) Other Sources: To further broaden the scope of the model's medical understanding, we included an additional 32.37B billion tokens from various supplementary sources, such as Articles, Medical Wikipedia, textbooks, and medical news articles[4]. These resources provide general medical knowledge, as well as updates on recent developments in healthcare and medical research.

To avoid the issue of catastrophic forgetting, commonly discussed in previous research,[19] we incorporated a subset of general-domain data alongside the medical domain corpus. For this purpose, we utilized the Fineweb dataset,[30] which comprises more than 15 trillion tokens of clean, deduplicated English web data sourced from CommonCrawl. We maintained a balance by setting the ratio of medical to general-domain data at approximately 4:1. As a result, we sampled 14.5 billion tokens from the general-domain corpus to complement the 57.97 billion tokens from the medical domain. We utilized Data-Juicer[31] to clean the dataset, eliminating overlaps and standardizing characters to ensure consistent Unicode formatting.

Table 1. Overview of continual pre-training datasets.

Data	Source	Size
Pile-Pubmed-Article	PubMed Central	20.86B
Pile-Pubmed-Abstract	PubMed	4.28B
Pile-NIH-Grant	NIH Reporter	358M
Healix-Shot	Wikipedia, News et al	32.38B
Guidelines	Clinical guidelines	999.7M
Fineweb	CommonCrawl	14.5B
Total	-	72.47B

2.1.2 Training Details

For the development of Lite-Me-LLaMA, we used LLaMA3-8B as the backbone model. The pre-training process followed the standard language modeling objective,[?] aimed at maximizing the likelihood of token sequences within the training data. The objective function is defined as: $\mathcal{L}(\Theta) = \sum_i^k \log P_\Theta(x_i|x_1, x_2, \cdots, x_{i-1})$, where $x = \{x_1, x_2, \cdots, x_k\}$ is a given input sequence of tokens, and Θ is the parameters of LLaMA3-8B model. The model was trained for one epoch using five nodes with 8 NVIDIA A100 80G GPUs from the super computing cluster at University of Florida, a process that took approximately 110 hours. The maximum sequence length was set to 8192 tokens. The training configuration included a learning rate of 1e-5, weight decay of 0.00001, and a warmup ratio of 0.05. Additionally, we employed bf16 precision, with a seed value of 1234. For optimization, we set gradient accumulation steps to 16, with a per-device batch size of 3. The training process utilized DeepSpeed [5] to enhance the efficiency of large-scale model training.

2.2 Lite-Me-LLaMA-Instruct: Adapt Lite-Me-LLaMA for Instruction Following by Instruction Fine-tuning

To further improve the model's ability to follow instructions and directly perform on new medical tasks without additional training, we developed Lite-Me-LLaMA-Instruct-8B by fine-tuning Lite-Me-LLaMA using a comprehensive medical instruction data.

Table 2. Overview of instruction datasets. Data marked with "*" indicates novel data created by us.

Data	Task	Source	Size
Anki flashcards[32]	Open-ended QA	Online learning tool	34K
Wikidoc Patient[32]	Open-ended QA	Wikidoc	5.94K
Wikidoc Textbook[32]	Open-ended QA	Wikidoc	10K
MediQA[33]	Open-ended QA	Online QA system	2.21k
LiveQA[34]	Open-ended QA	Online QA system	529
MedicationQA[35]	Open-ended QA	Online QA system	690
MedicalQA[36]	Open-ended QA	Clinical trials, PubMed	426K
GuidelineQA[19]	Open-ended QA	Clinical Guidelines	2K
MMLU medical[37]	Multiple-choice QA	Online sources	3.79K
Genetic QA*	Open-ended QA	Medical book	9.54K
Conversations[38]	Real conversations	Online forums	50K
MedInstruct[39]	Open-source instructions	ChatGPT generated	52K
Pubmed Case*	Diagnosis prediction	PubMed	33.38K
Dolly[40]	Open-source instructions	human generated	15K
Total after filtering	-	-	602K

Table 3. Overview of evaluation datasets.

Data	Train	Test	Choices
PubMedQA	211,269	500	3
MedQA	10,178	1,273	5
MedMCQA	182,822	4,183	4

2.2.1 Medical Instruction Dataset

We compiled a diverse and comprehensive instruction-tuning dataset, specifically curated to enhance Lite-Me-LLaMA's instruction-following capabilities for medical applications, as detailed in Table 2. This dataset comprises 602K samples, sourced from a range of publicly available datasets and novel datasets created for this project. The dataset integrates several key sources: (1) Anki flashcards,[32] which include 34K open-ended question-answer pairs derived from an online learning

[4] https://huggingface.co/datasets/health360/Healix-Shot

[5] https://github.com/microsoft/DeepSpeed

Table 4. The supervised fine-tuning performance of different models.

Model	PubMedQA		MedQA		MedMCQA	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
LLaMA2 7B	0.758	0.529	0.428	0.425	0.503	0.496
Meditron 7B	0.758	0.452	0.462	0.459	0.555	0.549
LLaMA2 13B	0.800	0.560	0.467	0.465	0.527	0.524
PMCLLaMA 13B	0.778	0.544	0.456	0.454	0.548	0.545
Me-LLaMA 13B	**0.802**	**0.562**	0.493	0.487	0.557	0.551
LLaMA3 8B	0.750	0.519	0.496	0.497	0.574	0.563
Lite-Me-LLaMA 8B	0.777	0.538	**0.539**	**0.533**	**0.583**	**0.575**

tool; (2) Wikidoc Patient and Textbook datasets,[32] contributing 15.94K combined samples of medical questions and answers from Wikidoc; (3) MediQA,[33] with 2.21K open-ended QA pairs from an online medical QA system; and (4) MedicalQA,[36] a large dataset of 426K open-ended QA examples based on clinical trials and PubMed articles. In addition, we included a variety of specialized data: (5) GuidelineQA,[19] featuring 2K QA pairs drawn from clinical guidelines; (6) Genetic QA, a new dataset we developed, containing 9.54K open-ended QA pairs based on a genetic review medical textbook [6]; and (7) Pubmed Case, another novel dataset created by us, with 33.38K samples focused on diagnosis prediction of patient cases sourced from PubMed. We also incorporated conversational data (50K samples) from online medical forums[38] to enhance the model's ability to manage real-world interactions, and MedInstruct,[39] an open-source dataset of 52K medical instruction-response pairs generated via ChatGPT [7]. We finally mixed a general domain data Dolly,[40] with 15K human-generated open-source instruction-response pairs. To maintain data quality, we conducted extensive filtering and deduplication processes. We removed instruction pairs where either the input or output was null, ensuring that only complete, meaningful interactions were included. For open-ended datasets, we applied additional filters by removing samples where the output was fewer than 20 tokens or where the input contained fewer than 20 tokens. After these efforts, the final dataset includes 602K samples. The prompt of each dataset is shown in Appendix A, Table 6.

2.2.2 Training Details

We fine-tuned Lite-Me-LLaMA for 1 epoch. The model was optimized on 1 NVIDIA A100 40GB GPUs, with the entire training process taking approximately 48 hours. We utilized a learning rate of 1e-4 and applied a weight decay of 0.00001 and a warmup ratio of 0.01 to regularize the model during training. The fine-tuning process employed gradient accumulation with a step size of 1. We utilized parameter-efficient tuning with LoRA[41] settings of $r = 64$, $\alpha = 128$ and no dropout.

[6] https://www.ncbi.nlm.nih.gov/books/NBK1116

[7] https://openai.com/chatgpt/

3 EXPERIMENTS

3.1 Evaluation Datasets

To assess the performance of Lite-Me-LLaMA, we followed established evaluation protocols from previous studies[19,42] and used three widely-used medical question-answering (QA) datasets: (1) PubMedQA[43]: Given a PubMed abstract and a question, the model predicts one of three answers: yes, no, or maybe. We use 500 test samples following existing studies. (2) MedQA[44]: This dataset contains questions styled after the US Medical License Exam (USMLE). It tests a broad range of medical knowledge, including patient profiles, symptoms, and treatments. We use the subset with five answer options, which includes 10,178 training samples and 1,273 test samples. (3) MedMCQA[45]: A dataset comprising over 194K multiple-choice questions with four answer options, sourced from Indian medical entrance exams. It covers 21 medical subjects and 2.4K healthcare topics. We evaluate using the 4,183-sample validation set, as the test set does not have publicly available answers. As shown in Appendix B, Table 7, the standardized prompts used for each test dataset to ensure consistency in the evaluation.

3.2 Evaluation Setting

Supervised Learning. In the supervised learning setting, we fine-tuned Lite-Me-LLaMA using the training sets of the evaluation datasets independently outlined earlier. We used the AdamW optimizer[46] for training up to 3 epochs. Across all datasets, a consistent learning rate of 1e-4 was applied. After fine-tuning, the performance of the models was evaluated on the corresponding test sets, allowing us to measure their task-specific capabilities. For comparison, we included general-purpose lightweight LLMs (models with fewer than 13B parameters), such as LLaMA2 7B/13B,[47] LLaMA3 8B,[48] as well as medical-specific lightweight LLMs, including

Table 5. The zero-shot learning performance of different models.

Model	PubMedQA		MedQA		MedMCQA	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
LLaMA2-7B-chat	0.572	0.288	0.214	0.070	0.322	0.121
LLaMA2-13B-chat	0.546	0.457	0.097	0.148	0.321	0.243
PMCLLaMA-13B-chat	0.504	0.305	0.207	0.158	0.212	0.216
Medalpaca-13B	0.238	0.192	0.143	0.102	0.205	0.164
AlpaCare-13B	0.538	0.373	0.304	0.281	0.385	0.358
Me-LLaMA-13B-chat	0.700	**0.504**	0.427	0.422	0.449	0.440
LLaMA3-8B-Instruct	0.650	0.420	0.487	0.405	0.459	0.401
Lite-Me-LLaMA-Instruct	**0.704**	0.500	**0.508**	**0.501**	**0.487**	**0.482**

Meditron 7B,[18] PMC-LLaMA 13B,[42] and Me-LLaMA 13B.[19]

Zero-shot. For the zero-shot learning evaluation, we tested Lite-Me-LLaMA-Instruct's ability to handle unseen tasks without any task-specific fine-tuning. We compared Lite-Me-LLaMA-Instruct's zero-shot performance against several baseline models that excel in instruction following. These baselines included general purpose LLMs LLaMA2-7B/13B-chat,[47] LLaMA3-8B-Instruct,[48] and instruction-tuned medical LLMs including PMCLLaMA-13B-chat,[42] Medalpaca-13B,[32] AlpaCare-13B,[39] and Me-LLaMA-13B-chat.[19]

3.3 Performance

3.3.1 Supervised Learning Performance

As shown in Table 4, Lite-Me-LLaMA 8B demonstrates clear superiority over all baseline models in the supervised learning setting, including those with significantly larger parameter sizes. In particular, Lite-Me-LLaMA 8B surpasses models such as LLaMA2 13B and PMC-LLaMA 13B on the MedQA benchmark, highlighting the efficiency of smaller models equipped with domain-specific knowledge. This performance illustrates that continual pre-training on specialized medical corpora can bridge the gap typically attributed to model size, allowing smaller models to outperform their larger counterparts. Furthermore, Lite-Me-LLaMA 8B outperforms its backbone model, LLaMA3 8B, despite sharing the same architecture, underscoring the critical impact of continual pre-training in adapting general-purpose models to excel in domain-specific tasks. These findings emphasize that specialized pre-training, rather than sheer model size, can be the determining factor in achieving superior performance in medical AI applications.

3.3.2 Zero-shot Learning Performance

In the zero-shot learning evaluation, Lite-Me-LLaMA-Instruct exhibited outstanding generalization capabilities across unseen medical tasks without any task-specific fine-tuning, as shown in Table 5. Compared to both general-purpose instruction-following models and other medical LLMs, Lite-Me-LLaMA-Instruct consistently achieved superior performance, particularly against larger models like PMC-LLaMA-13B-chat, MedAlpaca-13B, AlpaCare-13B, and Me-LLaMA-13B-chat. Despite its smaller size, Lite-Me-LLaMA-Instruct outperformed these models across various benchmarks, underscoring the value of domain-specific continual pre-training. This pattern holds true when comparing Lite-Me-LLaMA-Instruct to its backbone model, LLaMA3-8B-Instruct, with substantial gains in performance attributed to its continual pre-training on medical datasets and instruction tuning. These findings emphasize the impact of domain adaptation in boosting zero-shot capabilities, highlighting the efficiency and power of tailored pre-training for medical tasks, even in resource-constrained environments.

4 DISCUSSION

Our experimental findings highlight that Lite-Me-LLaMA outperforms both lightweight open-source medical LLMs (around 8B parameters) and larger models (13B parameters), while maintaining low resource demands, making it an ideal tool for clinicians and healthcare institutions. Designed for local deployment with minimal computational infrastructure, Lite-Me-LLaMA provides a cost-effective solution, allowing hospitals and clinics to implement advanced AI models without the need for large-scale computing resources. The Lite-Me-LLaMA base model can be easily fine-tuned with task-specific datasets using just one NVIDIA A100 40GB GPU, enabling healthcare providers to tailor the model to their unique needs, such as diagnostic support, patient record analysis, or clinical decision-making. Clinicians can directly use Lite-Me-LLaMA-Instruct for real-time tasks through prompt engineering, offering immediate support for various medical tasks without extensive retraining. This makes the model highly adaptable to clinical workflows and significantly reduces deployment complexity, even in resource-constrained environments where multiple consumer-grade GPUs can be used in-

stead of high-end hardware.

One of the key advantages of Lite-Me-LLaMA is its ability to ensure data privacy and transparency. Since the models are locally deployable, sensitive patient data can be processed in-house, eliminating the need for external servers and reducing the risk of data breaches or non-compliance with privacy regulations such as HIPAA. Furthermore, because Lite-Me-LLaMA is open-source, clinicians and institutions gain full visibility into how the model operates, offering transparency that is often lacking in closed-source, proprietary models. This transparency fosters trust in the model's outputs, ensuring that healthcare providers can rely on the AI to support critical medical decisions.

Despite its robust capabilities, Lite-Me-LLaMA does have limitations, particularly in its reliance on the LLaMA3-8B backbone, which may not be as efficient as newer iterations like LLaMA3.1 and LLaMA3.2. Additionally, the current model handles only textual data, and expanding it to support multimodal inputs like medical images and charts would significantly increase its applicability in comprehensive clinical workflows. Looking ahead, we are committed to the continual development of lightweight, open-source medical LLMs by integrating more advanced backbone models and exploring multimodal capabilities. This ongoing work aims to enhance efficiency, performance, and versatility, ensuring that Lite-Me-LLaMA remains a powerful, accessible tool for healthcare providers.

COMPETING INTERESTS

No competing interest is declared.

AUTHOR CONTRIBUTIONS STATEMENT

REFERENCES

1. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. Jama. 2005;293(10):1223-38.

2. Jaspers MW, Smeulers M, Vermeulen H, Peute LW. Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. Journal of the American Medical Informatics Association. 2011;18(3):327-34.

3. Lighthall GK, Vazquez-Guillamet C. Understanding decision making in critical care. Clinical medicine & research. 2015;13(3-4):156-68.

4. Ely JW, Osheroff JA, Chambliss ML, Ebell MH, Rosenbaum ME. Answering physicians' clinical questions: obstacles and potential solutions. Journal of the American Medical Informatics Association. 2005;12(2):217-24.

5. Yang Q, Hao Y, Quan K, Yang S, Zhao Y, Kuleshov V, et al. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023. p. 1-14.

6. Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. JAMA internal medicine. 2014;174(5):710-8.

7. Daei A, Soleymani MR, Ashrafi-Rizi H, Zargham-Boroujeni A, Kelishadi R. Clinical information seeking behavior of physicians: A systematic review. International journal of medical informatics. 2020;139:104144.

8. Tomašev N, Harris N, Baur S, Mottram A, Glorot X, Rae JW, et al. Use of deep learning to develop continuous-risk models for adverse event prediction from electronic health records. Nature Protocols. 2021;16(6):2765-87.

9. Lee M, Cimino J, Zhu HR, Sable C, Shanker V, Ely J, et al. Beyond information retrieval—medical question answering. In: AMIA annual symposium proceedings. vol. 2006. American Medical Informatics Association; 2006. p. 469.

10. Kell G, Marshall IJ, Wallace BC, Jaun A. What would it take to get biomedical QA systems into practice? In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. vol. 2021. NIH Public Access; 2021. p. 28.

11. Abacha AB, Shivade C, 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; 2019. p. 370-9.

12. Lakkaraju H, Slack D, Chen Y, Tan C, Singh S. Rethinking explainability as a dialogue: A practitioner's perspective. arXiv preprint arXiv:220201875. 2022.

13. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nature medicine. 2023;29(8):1930-40.

14. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, et al. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

15. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259-65.

16. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine. 2023;183(6):589-96.

17. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, Hogarth M, et al. Evaluating artificial intelligence responses to public health questions. JAMA network open. 2023;6(6):e2317517-7.

18. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al. Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:231116079. 2023.

19. Xie Q, Chen Q, Chen A, Peng C, Hu Y, Lin F, et al. Me llama: Foundation large language models for medical applications. arXiv preprint arXiv:240212749. 2024.

20. Toma A, Senkaiahliyan S, Lawler PR, Rubin B, Wang B. Generative AI could revolutionize health care—but not if control is ceded to big tech. Nature. 2023;624(7990):36-8.

21. Saab K, Tu T, Weng WH, Tanno R, Stutz D, Wulczyn E, et al. Capabilities of gemini models in medicine. arXiv preprint arXiv:240418416. 2024.

22. Irugalbandara C, Mahendra A, Daynauth R, Arachchige TK, Flautner K, Tang L, et al. Scaling Down to Scale Up: A Cost-Benefit Analysis of Replacing OpenAI's GPT-4 with Self-Hosted Open Source SLMs in Production. arXiv preprint arXiv:231214972. 2023.

23. Umeton R, Kwok A, Maurya R, Leco D, Lenane N, Willcox J, et al. GPT-4 in a cancer center—institute-wide deployment challenges and lessons learned. NEJM AI. 2024;1(4):AIcs2300191.

24. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. The Lancet Digital Health. 2023;5(6):e333-5.

25. Chen X, Mao X, Guo Q, Wang L, Zhang S, Chen T. RareBench: Can LLMs Serve as Rare Diseases Specialists? In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery; 2024. p. 4850–4861.

26. Chen L, Zaharia M, Zou J. Frugalgpt: How to use large language models while reducing cost and improving performance. arXiv preprint arXiv:230505176. 2023.

27. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. Qlora: Efficient finetuning of quantized llms. Advances in Neural Information Processing Systems. 2024;36.

28. Peng C, Yang X, Chen A, Smith KE, PourNejatian N, Costa AB, et al. A study of generative large language model for medical research and healthcare. NPJ digital medicine. 2023;6(1):210.

29. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:210100027. 2020.

30. Penedo G, Kydlíček H, allal LB, Lozhkov A, Mitchell M, Raffel C, et al.. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale; 2024.

31. Chen D, Huang Y, Ma Z, Chen H, Pan X, Ge C, et al. Data-juicer: A one-stop data processing system for large language models. In: Companion of the 2024 International Conference on Management of Data; 2024. p. 120-34.

32. Han T, Adams LC, Papaioannou JM, Grundmann P, Oberhauser T, Löser A, et al. MedAlpaca–an open-source collection of medical conversational AI models and training data. arXiv preprint arXiv:230408247. 2023.

33. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Scientific Data. 2020;7(1):322.

34. Abacha AB, Agichtein E, Pinter Y, Demner-Fushman D. Overview of the medical question answering task at TREC 2017 LiveQA. In: TREC; 2017. p. 1-12.

35. Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the gap between consumers' medication questions and trusted answers. In: MEDINFO 2019: Health and Wellbeing e-Networks for All. IOS Press; 2019. p. 25-9.

36. Vsevolodovna RM. AI Medical Dataset; 2023. Available from: `https://github.com/ruslanmv/ai-medical-chatbot`.

37. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring Massive Multitask Language Understanding. In: International Conference on Learning Representations;. .

38. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. Cureus. 2023;15(6).

39. Zhang X, Tian C, Yang X, Chen L, Li Z, Petzold LR. Alpacare: Instruction-tuned large language models for medical application. arXiv preprint arXiv:231014558. 2023.

40. Conover M, Hayes M, Mathur A, Xie J, Wan J, Shah S, et al.. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM; 2023.

41. Hu EJ, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, et al. LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations;. .

42. Wu C, Lin W, Zhang X, Zhang Y, Xie W, Wang Y. PMC-LLaMA: toward building open-source language models for medicine. Journal of the American Medical Informatics Association. 2024:ocae045.

43. Jin Q, Dhingra B, Liu Z, Cohen 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); 2019. p. 2567-77.

44. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021;11(14):6421.

45. Pal A, Umapathi LK, Sankarasubbu M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on health, inference, and learning. PMLR; 2022. p. 248-60.

46. Loshchilov I. Decoupled weight decay regularization. arXiv preprint arXiv:171105101. 2017.

47. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:230709288. 2023.

48. Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, et al. The llama 3 herd of models. arXiv preprint arXiv:240721783. 2024.

A MEDICAL INSTRUCTION DATA

Table 6 shows the prompt of each dataset for instruction fine-tuning the Lite-Me-LLaMA model.

Table 6. The prompt of each instruction fine-tuning dataset.

Data	Prompt
Anki Flashcards	"If you are a medical professional, answer this question truthfully. INPUT: {text} OUTPUT: "
Wikidoc Patient	"Answer this medical question truthfully. INPUT: {text} OUTPUT: "
Wikidoc Textbook	"Answer this medical question truthfully. INPUT: {text} OUTPUT: "
MediQA	"Answer the input medical question based on the given context. INPUT: {text} CONTEXT: {text} OUTPUT: "
LiveQA	"Given a medical query, provide a concise and clear answer based on the given details. INPUT: {text} OUTPUT: "
MedicationQA	"Answer this medical question truthfully. INPUT: {text} OUTPUT: "
GuidelineQA	"If you are a medical professional, answer this question truthfully. INPUT: {text} OUTPUT: "
MMLU Medical	"Given a medical question and four options, select the correct answer from the four options. INPUT: {text} OUTPUT: "
Genetic QA	"You are an expert in genetics and gene-related diseases. Give a question related to a specific gene disease, your task is to answer the questions accurately and concisely. INPUT: {text} OUTPUT: "
Conversations	"Given a medical query, provide a concise and clear answer based on the patient's description. INPUT: {text} OUTPUT: "
MedInstruct	"Below is an input that describes a medical task, maybe paired with a context that provides further input information. Write a response that appropriately completes the request. INPUT: {text} CONTEXT: {text} OUTPUT: "
Pubmed Case	"As a physician engaged in clinical diagnostic reasoning, here is a case report from Pubmed, please analyze the input and accurately identify the diagnosis presented within it. INPUT: {text} OUTPUT: "
Dolly	"Below is an input that describes a task, maybe paired with a context that provides further information. Write a response that appropriately completes the request. INPUT:{Text} CONTEXT:{Text} OUTPUT: "

Table 7. The prompt of each QA test dataset.

Data	Prompt
PubMedQA	"TASK: Your task is to answer the biomedical questions based on the provided input. Only output yes, no, or maybe as answer. INPUT:{Text} Question:{Text} OUTPUT:"
MedQA	"You are a medical doctor taking the US Medical Licensing Examination. You need to demonstrate your understanding of basic and clinical science, medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy. Show your ability to apply the knowledge essential for medical practice. For the following multiple-choice question, select one correct answer from A to E. Base your answer on the current and standard practices referenced in medical guidelines. Question:{text} Options: {text} Answer:"
MedMCQA	"You are a medical doctor taking the US Medical Licensing Examination. You need to demonstrate your understanding of basic and clinical science, medical knowledge, and mechanisms underlying health, disease, patient care, and modes of therapy. Show your ability to apply the knowledge essential for medical practice. For the following multiple-choice question, select one correct answer from A to D. Base your answer on the current and standard practices referenced in medical guidelines. Question:{text} Options: {text} Answer:"

B EXPERIMENTS

Table 7 shows the prompt of each QA test dataset for evaluation.