Benchmarking Large-Language Models for Resource-Efficient Medical AI for Edge Deployment

Awal Ahmed Fime[1,2,*], Md Zarif Hossain[1,2,*], Saika Zaman[1], Abdur Rahman Bin Shahid[1], Ahmed Imteaj[1,2,*]

[1]School of Computing, Southern Illinois University Carbondale, IL, USA
[2]Security, Privacy and Intelligence for Edge Devices Laboratory (SPEED Lab)
awalahmed.fime@siu.edu, mdzarif.hossain@siu.edu, saika.zaman@siu.edu, shahid@cs.siu.edu, imteaj@cs.siu.edu

Abstract

Large-Language Models (LLMs) are rapidly emerging as transformative tools across diverse domains, leveraging extensive training on vast and heterogeneous datasets to capture nuanced knowledge and transcend traditional boundaries of understanding. In the medical domain, LLMs hold immense potential to revolutionize clinical workflows by enhancing the efficiency of medical practitioners and alleviating their workload. However, a critical gap exists between the theoretical capabilities of LLMs and their practical deployment in resource-constrained environments, such as edge devices (e.g., health monitors) commonly used in healthcare settings. This paper addresses this challenge by employing parameter-efficient fine-tuning (PEFT) techniques to adapt widely available advanced LLMs for medical applications while comparing their resource efficiency and performance. The models are fine-tuned on structured medical question answering datasets, and their outputs are evaluated using BERTScore and USEScore metrics. Among the models tested, Mistral v0.3 demonstrated the best performance based on both metrics, while also exhibiting promise for resource efficiency. These findings provide a vital foundation for selecting and optimizing LLMs for healthcare tasks, offering actionable insights for developing resource-efficient and scalable solutions that are well-suited for deployment on edge devices in real-world medical environments.

Introduction

In recent years, Large-Language Models (LLMs) have increasingly demonstrated their ability to interact with humans in a manner that is both coherent and contextually appropriate. When posed with a question, LLMs can interpret the context and generate a response that aligns with the intent and content of the inquiry. One of the remarkable features of LLMs is their ability to provide answers even to questions they have never encountered before, leveraging their extensive training on contextually similar data. This capability underscores the importance of training LLMs on large and varied datasets to ensure they can generalize across different scenarios and domains. Integrating LLMs into healthcare applications can optimize medical workflows, leveraging generative AI to produce contextually appropriate and medically sound responses. This shift not only supports healthcare professionals but also benefits patients, enabling advanced deployment of AI-powered tools for better decision support. However, efficient training and inference of LLMs for medical applications pose several challenges, particularly in terms of optimizing memory and computational resources. Medical datasets can be large, complex, and resource-intensive, making it imperative to balance model performance with computational efficiency. As such, the deployment of LLMs in healthcare, particularly on edge devices, requires careful consideration of hardware limitations and the need for real-time responses in clinical environments. Numerous studies have demonstrated the effectiveness of LLMs in handling both generic (Liu et al. 2023; Biderman et al. 2023) and domain-specific [e.g., Governance (Mamalis et al. 2024), Finance (Lakkaraju et al. 2023), Law (Rafat 2024)] tasks. Comparing and benchmarking various LLMs is a common practice, providing insights into their performance across different areas (Zheng et al. 2023). However, benchmarking LLMs specifically for the medical domain remains limited. While some datasets (Han et al. 2023; Li et al. 2023), and studies (Chen et al. 2024; Anil et al. 2023) focus on applying LLMs to medical tasks, there is a lack of studies that comprehensively compare both the performance of these models and their resource-efficiency in medical domain. This highlights the need for structured evaluations to better understand their applicability in healthcare. To bridge this gap, we conducted a comprehensive comparison of several state-of-the-art LLMs. Using parameter-efficient fine-tuning (PEFT), we adapted these models to medical applications and evaluated their performance across five distinct healthcare-related datasets. The evaluation leveraged two key metrics, BERTScore and USEScore, to provide a robust assessment of model effectiveness. The results of this analysis provide valuable insight into the strengths and limitations of these models, helping to identify the most suitable LLM for medical tasks.

Literature Review

LLMs have drawn significant interest in the medical-related domain in recent times, particularly for medical question-answering tasks. Various strategies have been introduced to improve the accuracy, resource efficiency, and usability of LLMs. For instance, ChatDoctor (Li et al. 2023) is

*Corresponding author.

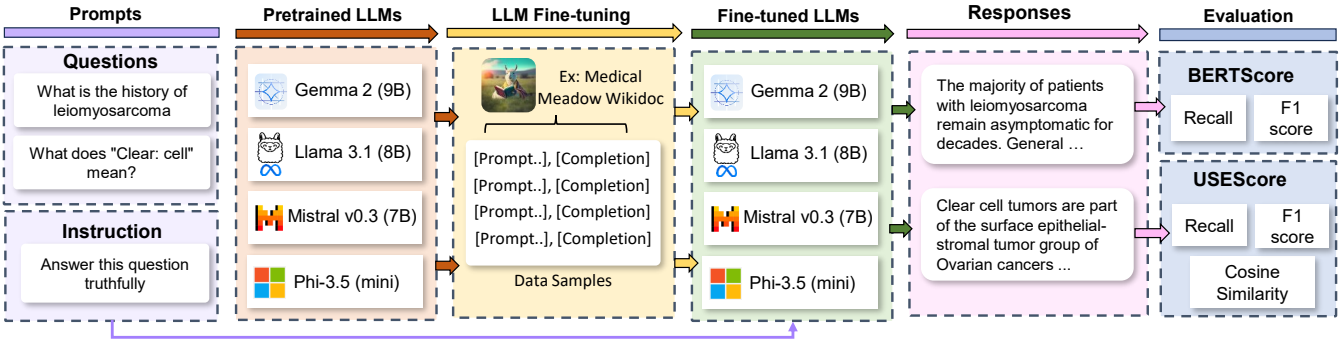

Figure 1: Evaluation workflow of LLMs for Medical Question-Answering considering four LLMs — Gemma 2 (9B), Llama 3.1 (8B), Mistral v0.3 (7B), and Phi-3.5 (Mini) — on diverse medical datasets. Performance is measured with BERTScore and USEScore metrics, focusing on Recall, F1 score, and Cosine Similarity.

a fine-tuned version of the LLaMA model with 7B parameters, specifically tuned for medical question-answering tasks. This model was pre-trained on the Alpaca dataset (Taori et al. 2023) and later fine-tuned using a dataset of $100,000$ patient conversations to enhance its performance in the medical domain. Similarly, Chen et al. (Chen et al. 2024) examined the impact of fine-tuning LLMs for the medical domain, investigating both the training data and model architecture. In addition to these models, Med-PaLM (Singhal et al. 2023) stands out as a domain-specific adaptation of PaLM, leveraging few-shot, chain-of-thought (CoT), and self-consistency prompting strategies to tailor the general-purpose PaLM model (Chowdhery et al. 2023) for medical question answering. By utilizing six publicly available datasets and further refining the model with a prompt based on 40 examples, Med-PaLM achieved state-of-the-art performance in this specialized task. PaLM-2 (Anil et al. 2023), an enhanced version of PaLM, introduced an ensemble refinement (ER) prompting strategy. This two-stage approach first generates multiple candidate responses using few-shot prompts, followed by a second stage where the model reconsiders the original prompt to explain and refine its answer. Utilizing the mixture-of-expert (MOE) for multi-task learning and Low-Rank Adaptation (LoRA), (Liu et al. 2024) presented a layer termed MOELoRA for multi-task medical application. To control the contribution of MOELoRA they proposed a task-motivated gate function.

Methodology

Pipeline

This segment describes the workflow of LLMs for a medical question-answering task Figure 1. In this study, four different LLMs were used: Gemma 2 (9B), Llama 3.1 (8B), Mistral v0.3 (7B), and Phi-3.5 (mini). Each model is trained using the same hyperparameters to maintain consistency throughout the experiments. These hyperparameters includes learning rate, batch size, and number of training epochs, allowing for a fair comparison of model performance. The training phase involved exposing the models to the prompts to optimize their ability to generate relevant and accurate responses. During the inference phase, the generated responses are compared to reference outputs, which are typically ground truth answers from the training set, to assess the models' accuracy and relevance in responding to medical queries. For evaluation, two metrics are used to measure the performance of the models. First, BERTScore assesses the semantic similarity between the generated responses and the reference outputs. This metric evaluates the degree to which the generated response aligns with the meaning of the reference by comparing word embeddings, capturing nuances of meaning even if the wording differs. Second, USEScore (Universal Sentence Encoder Score) is utilized to measure sentence-level semantic similarity. USEScore uses sentence embeddings generated by the Universal Sentence Encoder which provides a more holistic evaluation of how closely the entire generated sentence matches the reference in meaning.

Base Models

For our fine-tuning process, we leverage a diverse set of pretrained LLMs, selected for their widespread recognition in the field and publicly available source code. Gemma (Team et al. 2024), a family of lightweight models by Google derived from the Gemini architecture, excels in English text generation tasks like question answering and summarization on low-resource hardware. Llama 3.1 (8B) (Dubey et al. 2024), an evolution of Meta's Llama (Touvron et al. 2023), incorporates supervised fine-tuning and reinforcement learning with human feedback (RLHF). It supports eight languages and was trained on over 15T tokens and 25M synthetic examples with 39.3M GPU hours of computation. Mistral v0.3 (7B) (Jiang et al. 2023), utilizing grouped-query attention and sliding window attention, specializes in handling long-range dependencies and sequences efficiently, outperforming larger Llama models in benchmarks while focusing on English. Lastly, Phi-3.5-mini (Abdin et al. 2024), a compact 3.8B-parameter model designed for reasoning-intensive tasks like math and logic, supports a 128k token context length and was trained on high-quality synthetic and public datasets across 23 languages.

Efficient Training

The process of optimizing LLM requires the intricate coordination of several parameters, each of which needs to

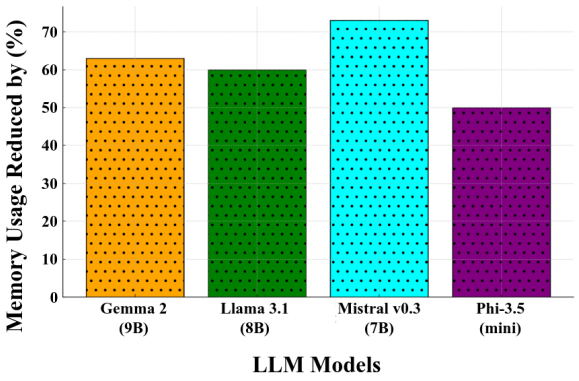

Figure 2: Efficient use of VRAM for training.

be precisely calibrated, saved, and processed using intricate mathematical procedures. Though computationally challenging, careful refinements can greatly increase the efficiency of this complex operation. It becomes possible to not just speed up training but also lower memory consumption by reconsidering how models handle these operations—whether through adaptive computing, strategic storing, or simplified parameter updates.

Reducing Training Time The training time of each of the four models [Phi-3.5 (Mini), Gemma 2 (9B), Llama 3.1 (8B), and Mistral v0.3 (7B)] varies depending on their architectural optimizations and model size. The smallest, Phi-3.5 (Mini), has fewer parameters to update by nature, which results in quicker training times. However, it lacks the advanced techniques, such as attention optimizations, typically employed in larger models. Even though Gemma 2 (9B) and Llama 3.1 (8B) have more parameters, they use more effective attention techniques, like multi-head attention with fused kernel optimizations, which dramatically reduce training-time matrix multiplications. By utilizing RoPE embeddings and improved tensor operations, Mistral v0.3 (7B), which attained the best size-performance trade-off, further minimizes training time. RoPE embeddings lessen the computing burden during backpropagation by enabling more effective attention computations across lengthy sequences. Moreover, Mistral v0.3 is built on an advanced causal mask technique, which speeds up the attention layer computation skipping unnecessary matrix reads that lead to an 8.1% reduction in training time, making it the most efficient among the models.

Memory Usage Optimization The memory usage of the models varies significantly due to their internal architecture and parameter sizes, as reflected in Figure 2. Phi-3.5 (Mini) has the smallest memory footprint, largely because of its reduced parameter count and simpler architecture. However, this limits its flexibility and performance on more complex tasks. On the other hand, Gemma 2 (9B) and Llama 3.1 (8B) manage their larger parameter sets with memory-efficient techniques such as mixed precision training and gradient checkpointing. These techniques reduce the need to store full-precision values for all operations, minimizing memory usage while retaining model accuracy. Mistral v0.3 (7B) goes a step further by employing parameter-efficient fine-tuning (PEFT) techniques, including LoRA and modular adapters, which freeze the majority of model parameters and only update task-specific components. This approach significantly reduces memory consumption during fine-tuning, as fewer parameters need to be stored and updated. Additionally, Mistral v0.3's architecture includes optimized memory management in attention layers, allowing for the efficient handling of long sequences without overwhelming memory resources, making it the most memory-efficient among the larger models.

Evaluation Metrics

To assess the performance of the models and their generated outputs, we employ robust and widely recognized evaluation metrics. These metrics are specifically designed to measure the semantic similarity and quality of generated sentences compared to reference sentences, providing insights into the effectiveness of the models. The metrics used are as follows:

BERTScore: BERTScore (Zhang et al. 2019) is an evaluation metric using contextual embeddings for text generation. It computes a similarity score for each candidate sentence with each token in the reference sentence. It uses pre-trained BERT (Kenton and Toutanova 2019) embeddings to convert words to tokens. Finally, it matches the similarity of the generated sentence and reference sentences. BERTScore computes precision, recall, and F1 scores from generated and reference sentences.

USEScore: Universal Sentence Encoder (USE) is a natural language processing technique that converts the text sentence to a higher dimension vector encoding (Cer et al. 2018). USE takes the variable length of input text and converts that to a vector of dimension 512. USE offers two types of encoding, one is a Transformer based another is a Deep Averaging Network. After encoding, the similarity of the metrics is calculated from generated sentences and reference sentences. The equation to calculate recall, precision and F1 score are shown in equation 1, 2 and 3:

$$R_{\text{USE}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{1}$$

$$P_{\text{USE}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \tag{2}$$

$$F_{\text{USE}} = 2 \cdot \frac{P_{\text{USE}} \cdot R_{\text{USE}}}{P_{\text{USE}} + R_{\text{USE}}} \tag{3}$$

Here, R_{USE} represents the calculated recall using USE encoding vector, where $\hat{x}$ and x are the vectors generated from the generated text and from the reference text, respectively. Similarly, P_{USE} and F_{USE} denote the precision and F1 score, respectively, both calculated using USE encoding vector.

Experimental Analysis

Datasets

Our experiments are conducted on five real-world medical question answering datasets: The Medical Meadow Wikidoc dataset (Han et al. 2023), which features question-

Attributes	Gemma	Llama	Mistral	Phi
Version	2	3.1	0.3	3.5
Parameters	9B	8B	7B	3.8
Trainable par.	54M	39M	42M	29M
Training Time	437.76	312.41	370.67	203.91

Table 1: Training Attribute Comparison of various LLMs.

answer pairs derived from WikiDoc. The AI Medical Chatbot (Vsevolodovna 2023) dataset is a large-scale database containing over 21.2M conversational texts between doctors and patients, collected from ClinicalTrials, EMEA, and PubMed. Lastly, the Mental Health dataset (Bertagnolli 2020), sourced from online counseling and therapy platforms, includes over 3.5k question-answer pairs in English, where qualified psychologists provide solutions. This dataset focuses on diverse mental health topics, enhancing LLMs' ability to provide mental health guidance. Additionally, we evaluate the LLMs on Medical Dataset (Jayantdocplix 2025), which contains over 3.7k data human and AI assistant conversion in English and the HealthCareMagic-100k-en dataset (Li et al. 2023), which includes over 112k medical records in English.

Experimental Setup

In this study, we compare four different models: Llama 3.1, Mistral v0.3, Phi-3.5, and Gemma. To ensure an unbiased comparison, we use models with the smallest number of parameters. All models are trained and tested with the same data distribution. We apply efficient fine-tuning techniques, such as LoRA, to enhance model performance. Rotary Position Embeddings Scaling is used with a maximum sequence length of 2048. For the training environment, two Tesla T4 GPUs are used. All models are trained for 2000 steps. The model is trained with learning rate 2e-5 and "adamw_8bit" optimizer. All models are trained with the same batch size 8.

LLM Fine-Tuning Process and Model Adaptation

We first initialize the pre-trained LLMs with weights derived from their training on extensive datasets. These weights provide the models with a comprehensive foundation of linguistic knowledge. Subsequently, the models are fine-tuned using question-answer pairings tailored for medical domains. We preprocess the datasets to ensure that each question and response is clear, consistent, and properly formatted.

LoRA Adaptation: LoRA utilizes low-rank matrices to approximate the parameter updates required for the model to effectively learn medical-specific language patterns and content, thereby avoiding the need to update the entire parameter set. LoRA lowers the computational and memory expenses while maintaining the model's ability to adjust to the subtleties of the medical questions and answers in the dataset by concentrating on the most important parameters, especially in the attention layers. With billions of parameters, large models like Llama and Mistral are especially benefiting from this strategy, which allows for effective fine-tuning without taxing the system's resources.

Fine-tune and Evaluation: In the fine-tuning stage, we optimized the LoRA parameters to generate accurate medical responses while keeping the majority of the pre-trained parameters fixed. Our fine-tuning procedure involved iterating over multiple epochs to minimize the loss function, which measured the error between the model's predictions and the expected responses. A summary of the base model architectures is presented in Table 1. Following fine-tuning, the models underwent a comprehensive evaluation to assess their effectiveness in the target tasks. This evaluation measured the models' semantic accuracy and contextual relevance using a range of metrics, including Recall, Precision, F1 Score, and Cosine Similarity. These metrics provided a detailed assessment of the models' ability to generate responses to medical queries that are both semantically accurate and contextually appropriate.

Results and Discussion

This section discusses the performance analysis of SOTA LLMs across diverse medical datasets and emphasizes their effectiveness in domain-specific applications. In Table 2, we provide a detailed comparative evaluation of the selected models, as discussed in the Methodology section. The evaluation was conducted on the test set of the baseline models trained on the five different medical datasets, offering critical insights into their relative strengths and limitations.

Performance Analysis Using BERTScore

Among the models, Phi-3.5 (mini), the smallest with only 3.8B parameters, exhibited the weakest performance, achieving an F1 score of 0.8182 on the Medical Meadow Wikidos and 0.7027 on the Health Care Magic dataset. It is consistently among the lowest-performing models across all other datasets. These results indicate that it struggles to maintain high semantic alignment with reference outputs, limiting its applicability in complex medical tasks. Llama, demonstrated significant improvement over Phi-3.5, with an F1 score of 0.8940 on the AI Medical Dataset. That is the highest among the results on this dataset. However, this model fails to maintain its consistency on other datasets. These results highlight its ability to leverage additional parameters effectively for improved performance, making it a more viable option for medical applications where accuracy is critical. Gemma 2 with 9B parameters further advanced precision, recall, and f1 metrics, outperforming Llama in most aspects. It achieved the highest accuracy on the AI Medical Dataset and maintained decent constancy over all other datasets. However, The top-performing model, Mistral v0.3 (7B), achieved the highest performance with an F1 score over 3 different datasets Medical Meadow Wikidos (0.8358), Medical Dataset (0.8893), and Health Care Magic (0.745). This result reflects a balanced and robust capability in generating semantically accurate outputs. Despite having fewer parameters than Gemma 2 and Llama 3.1, Mistral v0.3's architectural optimizations, including enhanced attention mechanisms, contribute to its superior performance. These results firmly establish Mistral v0.3 as the most effective model for medical tasks in this comparison.

Dataset	Model name	BERTScore			USEScore			
		Precision	Recall	F1	Cosine Sim.	Precision	Recall	F1
Medical Meadow Wikidoc (Han et al. 2023)	Gemma 2 (9B)	0.8321	0.8215	0.8259	0.4344	0.6031	0.6166	0.6098
	Llama 3.1 (8B)	0.8440	0.8181	0.8302	0.4237	0.6131	0.6281	0.6205
	Mistral v0.3 (7B)	0.8518	0.8217	0.8358	0.4531	0.6144	0.6401	0.6270
	Phi-3.5 (mini)	0.8177	0.8199	0.8182	0.4311	0.5999	0.6205	0.6101
Mental Health Dataset (Bertagnolli 2020)	Gemma 2 (9B)	0.8425	0.8332	0.8376	0.5403	0.6172	0.6795	0.6469
	Llama 3.1 (8B)	0.7593	0.8027	0.7796	0.2321	0.5358	0.4348	0.4800
	Mistral v0.3 (7B)	0.8402	0.8329	0.8363	0.5477	0.6256	0.6868	0.6547
	Phi-3.5 (mini)	0.8273	0.8249	0.8259	0.4847	0.5939	0.6408	0.6165
Medical Dataset (Jayantdocplix 2025)	Gemma 2 (9B)	0.8865	0.8816	0.8840	0.7961	0.8176	0.8103	0.8139
	Llama 3.1 (8B)	0.8855	0.8799	0.8826	0.7808	0.8049	0.8039	0.8044
	Mistral v0.3 (7B)	0.8908	0.8880	0.8893	0.8025	0.8197	0.8208	0.8202
	Phi-3.5 (mini)	0.8846	0.8871	0.8857	0.7976	0.8155	0.8178	0.8167
AI Medical Chatbot (Vsevolodovna 2023)	Gemma 2 (9B)	0.8759	0.8731	0.8744	0.6317	0.7642	0.7814	0.7727
	Llama 3.1 (8B)	0.9013	0.8878	0.8940	0.6663	0.7934	0.8006	0.7970
	Mistral v0.3 (7B)	0.8955	0.8883	0.8914	0.6596	0.7877	0.7891	0.7884
	Phi-3.5 (mini)	0.8702	0.8822	0.8756	0.6201	0.7704	0.7716	0.7710
Health Care Magic (Li et al. 2023)	Gemma 2 (9B)	0.7263	0.7638	0.7445	0.5947	0.8468	0.8422	0.8444
	Llama 3.1 (8B)	0.7215	0.7491	0.7359	0.5632	0.6598	0.7728	0.7117
	Mistral v0.3 (7B)	0.7274	0.7649	0.745	0.5954	0.8473	0.8429	0.8450
	Phi-3.5 (mini)	0.7039	0.7015	0.7027	0.5548	0.8285	0.8368	0.8325

Table 2: Comparative Evaluation of various LLMs on Diverse Medical Datasets using BERTScore and USEScore Metrics.

Performance Analysis Using USEScore

The evaluation using USEScore revealed a similar trend. Phi-3.5 (mini) again demonstrated the lowest performance, indicating substantial limitations in sentence-level semantic similarity. Conversely, Mistral v0.3 maintained its leading position, achieving the highest scores across four out of five datasets. This consistent performance across both evaluation frameworks underscores its well-rounded capability, excelling in both token-level and sentence-level assessments.

Critical Insights

This segment provides an in-depth analysis of the trade-offs between model size and semantic comprehension, precision-recall dynamics, architectural optimizations, and dataset-specific performance variations, highlighting the interplay of computational efficiency, domain adaptability, and task-specific strengths among the evaluated models.

Trade-Off Between Model Size and Semantic Comprehension: Phi (3.5) demonstrates the shortest training time (203.91 minutes) among all models, which aligns with its smaller parameter size (3.8B total parameters, 29M trainable parameters). This highlights Phi's computational efficiency but also its limited capacity for handling complex tasks requiring high representational power. On the other hand, Mistral v0.3 (7B) strikes a balance between parameter size, trainable parameters (42M), and training time (370.67 minutes), making it an optimal choice for tasks that require efficiency without significant compromises in performance.

Architectural Optimizations Drive Performance: The superior performance of Mistral v0.3, despite having fewer parameters than Gemma 2 and Llama 3.1, highlights the importance of architectural innovations. Features such as grouped-query attention and modular parameter-efficient fine-tuning strategies (e.g., LoRA) enable Mistral to outperform larger models. This suggests that efficiency-focused design can provide significant advantages in domains requiring both accuracy and computational scalability.

Model Efficiency and Training Requirements: Phi (3.5) demonstrates the shortest training time (203.91 minutes) and the smallest number of trainable parameters (29M), making it highly efficient in terms of computational resource usage. However, its smaller parameter size and reduced trainable parameters limit its ability to perform on complex datasets. Mistral v0.3 (7B), with a slightly larger training time (370.67 minutes) and 42M trainable parameters, strikes an effective balance, offering competitive performance while maintaining computational efficiency.

Dataset-Specific Performance Variations: While Mistral v0.3 performs exceptionally well overall, Llama 3.1 (8B) surpasses it on the AI Medical Chatbot Dataset, achieving the highest BERTScore F1 (0.9013) and USEScore F1 (0.8006), indicating its strength in conversational AI contexts. Conversely, Mistral v0.3 outperforms Llama models on structured datasets such as the Medical Meadow Wikidoc Dataset and the Health Care Magic Dataset, emphasizing its robustness in information-dense datasets.

Conclusion

This paper presents a comprehensive benchmarking of state-of-the-art LLMs for medical applications, addressing the critical need for evaluating their performance across diverse healthcare datasets. By fine-tuning and assessing multiple models using robust metrics such as BERTScore and USEScore, we provide a detailed comparison of their strengths, limitations, and suitability for specific medical tasks. The results underscore the importance of balancing computa-

tional efficiency, model size, and domain-specific adaptability when selecting models for healthcare-related use cases. Models like Mistral v0.3 demonstrated exceptional performance across structured datasets, while others, such as Llama 3.1, excelled in conversational contexts, highlighting the need for task-specific optimization. The findings also emphasize the potential for deploying efficient models like Mistral v0.3 on edge devices, enabling low-latency, resource-constrained applications in telemedicine and point-of-care scenarios. These insights pave the way for advancing AI-driven solutions in healthcare, with future work focusing on improving model robustness, expanding datasets, and optimizing architectures for edge-based deployments.

References

Abdin, M.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bertagnolli, N. 2020. Counsel chat: Bootstrapping high-quality therapy data.

Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.

Cer, D.; et al. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, 169–174.

Chen, J.; Jiang, Y.; Yang, D.; Li, M.; Wei, J.; Qian, Z.; and Zhang, L. 2024. Can LLMs' Tuning Methods Work in Medical Multimodal Domain? *arXiv preprint arXiv:2403.06407*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Han, T.; Adams, L. C.; Papaioannou, J.-M.; Grundmann, P.; Oberhauser, T.; Löser, A.; Truhn, D.; and Bressem, K. K. 2023. MedAlpaca–an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*.

Jayantdocplix. 2025. Medical Dataset Chat-Public Dataset in Hugging Face. https://huggingface.co/datasets/jayantdocplix/medical_dataset_chat. Accessed: 2025-01-25.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2.

Lakkaraju, K.; Jones, S. E.; Vuruma, S. K. R.; Pallagani, V.; Muppasani, B. C.; and Srivastava, B. 2023. LLMs for Financial Advisement: A Fairness and Efficacy Study in Personal Decision Making. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, 100–107.

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

Liu, Q.; Wu, X.; Zhao, X.; Zhu, Y.; Xu, D.; Tian, F.; and Zheng, Y. 2024. When MOE Meets LLMs: Parameter Efficient Fine-tuning for Multi-task Medical Applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1104–1114.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2023. GPT understands, too. *AI Open*.

Mamalis, M. E.; Kalampokis, E.; Fitsilis, F.; Theodorakopoulos, G.; and Tarabanis, K. 2024. A Large Language Model Agent Based Legal Assistant for Governance Applications. In *International Conference on Electronic Government*, 286–301. Springer.

Rafat, M. I. 2024. AI-powered Legal Virtual Assistant: Utilizing RAG-optimized LLM for Housing Dispute Resolution in Finland.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6): 7.

Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M. S.; Love, J.; et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vsevolodovna, R. M. 2023. AI Medical Dataset.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.