Llama-3[8B]-MeditronV1.0 — A State-of-the-Art Large Language Model for Medicine in 24 Hours

Alexandre Sallinen[1,2*], Guillaume Boye[1,2◕], Michael Zhang[1,2◕], Antoni-Joan Solergibert i Llaquet[1,2◕], Maud Dupont-Roc[1,3◕], Bastien Bernath [1,2◕], Etienne Boisson[1,2◕], Xavier Theimer-Lienhard[1,2◕], Hichem Hadhri[1,2◕], Antoine Tran[1,2◕], Veronique Suttels[5‡], Noémie Boillat-Blanco[5‡], Kristina Keitel[6‡], Mary-Anne Hartley[1,2,4*]

1 LiGHT: Laboratory for Intelligent Global Health and Humanitarian Response Technologies, Yale-EPFL
2 EPFL, School of Computer and Communication Sciences, Switzerland.
3 EPFL, School of Life Sciences, Switzerland.
4 Yale University, School of Medicine, Biomedical Informatics and Data Science, USA.
5 CHUV, Infectious Diseases Service, Rue du Bugnon 46, Lausanne, Switzerland.
6 Inselspital, Department of Pediatrics, Freiburgstrasse 20, Bern, Switzerland.

◕Meditron student team: Contributed equally.
‡Medical evaluation team: Contributed equally.

* Correspondence: mary-anne.hartley@yale.edu alexandre.sallinen@epfl.ch

Abstract

Introducing Llama-3[8B]-MeditronV1.0—a new medical Large Language Model (LLM) with 8 billion parameters, fine-tuned within 24-hours of the release of Meta's Llama-3. This new model outperforms all state-of-the-art open models within its parameter class on standard benchmarks such as MedQA and MedMCQA. It also outperforms Llama-2[70B] and is within 10% of the current leading open model for medicine in the 70B range: Llama-2[70B]-Meditron. This work shows the innovative potential of open foundation models with widely available weights and is part of a larger initiative to ensure equitable participatory access to this technology in low resource settings.

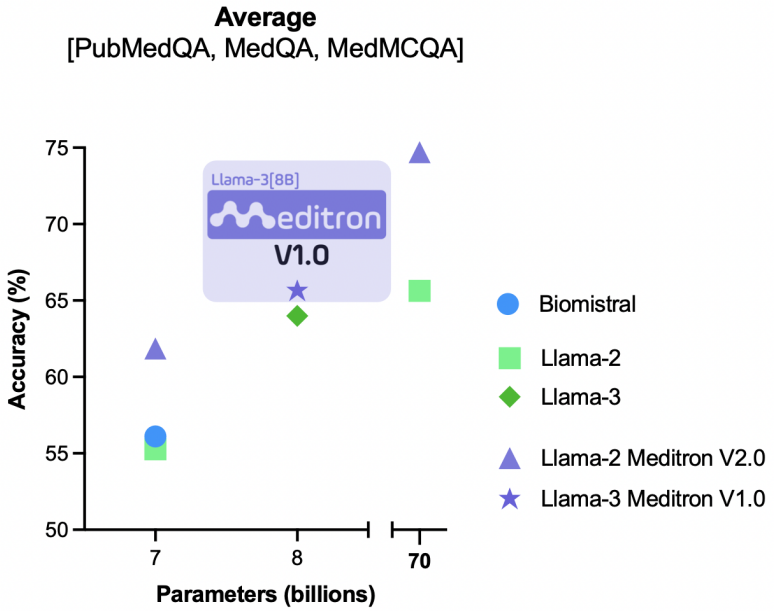


Fig 1. Average performance on MedQA, MedMCQA and PubmedQA according to parameter class

Introduction

The current leading language models (LLM) for medical applications (like Google's Med-PaLM 2), perform comparably to expert human physicians in high resource settings [1]. If appropriately adapted, such tools have the potential to offer lifesaving support in resource-constrained environments, where no physician is available. These closed-source, commercial tools do not represent the needs, populations or deployment realities of the settings in which they would have most impact. To respond to this issue, we previously developed Llama-2[70B]-Meditron: an open-source LLM for medicine, which is currently the state-of-the-art model across standard medical evaluation benchmarks. However, its large size poses deployment problems in low-resource settings, with large computational requirements and latency.

This work launched a fine tuning initiative and achieved SoTA results within its parameter class on the Llama3[8B] model within 24hours of its release.

Methods

Base model - Llama-3[8B] Llama3[2] is the latest generation open-source LLMs released by Meta on 18 April 2024. Llama-3[8B] is SOTA in its class diverse benchmarks. It employs a decoder-only transformer architecture and a 128K-token tokenizer, leveraging grouped query attention. The model is trained on over 15 trillion tokens from public sources and uses various filtering methods to maintain quality, including new safety tools and guidelines. The size of the dataset is 7 times that of Llama-2[3].

Dataset We finetune Llama-3[8B] prioritizing data quality over quantity. Our dataset is composed of a quality-filtered subset of PubMedCentral Public, which retains only meta analyses on human clinical studies, an open medical textbook set and our previously released diverse set of clinical practice guidelines [4].

How to fine tume Llama-3[8B] in less than 24 hours ? We optimized the training process to demonstrate the capabilities of our model when developed under a tight deadline. Speed-running the training on one epoch allowed us to quickly evaluate its performance and efficiency. After downloading the pre-trained Llama-3[8B] model, we employed the Axolotl framework[5] for fine-tuning and instruction-tuning, using packing for both. The choice of Axolotl guarantees no cross-contamination between the attentions.

Benchmarks We follow the evaluation procedure in Meditron[6] and use the three most popular benchmarks in medical LLM evaluation: PubMedQA[7], MedMCQA[8] and MedQA-4-Option [8]

- PubMedQA: This dataset presents a question-answering challenge using 'yes', 'no', or 'maybe' responses based on PubMed abstracts.
- MedMCQA: This dataset features over 194K questions from Indian medical entrance exams, used to create training and validation splits while omitting non-explanatory samples.
- MedQA-4-Option: MedQA is a dataset aiming at evaluating and fine-tuning medical LLMs with emphasis on chain-of-thought reasoning through human-written explanations.

Physician evaluation Similarly to [4], we asked 3 expert physicians to ask 10 challenging questions to Llama-3[8B]-Meditron and critically evaluate its answers based on a standard 12-point evaluation rubric. The adversarial questions were designed as a preliminary test of safety and bias.

Results

Llama-3[8B]-MeditronV1.0 is SOTA in its parameter class for standard benchmarks . Using the evaluation benchmarks PubMedQA, MedMCQA and MedQA, we compare our model (Llama-3[8B]-MeditronV1.0). Biomistral[7B][9], Llama-2[7B], Llama-2[70B], Llama-3[8B], Llama-2[7B]-Meditron Llama-2[70B]-Meditron.

We observe that our Llama-3[8B]-Meditron-V1.0 achieves SOTA results on 2 benchmarks in its 7B/8B category. On average, we beat all 7B and 8B models and achieve similar results to Llama-2-70B. Detailed results are in the appendix

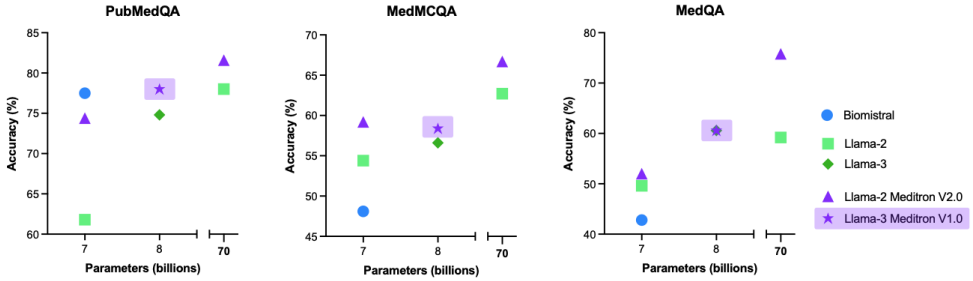


Fig 2. Performance on MedQA, MedMCQA and PubmedQA according to parameter class

Clinician evaluation Coming soon

Conclusion

Llama-3[8B]-MeditronV1.0 is a new state-of-the-art LLM for medicine in the 7/8B parameter class range. It was fine-tuned within 24-hours of the release of Llama-3, showing the potential of open foundation models with widely available wieghts.

Acknowledgments

We are thankful for the technical support provided by Antoine Bonnet and the medical advice of Léa Girani.

References

[1] Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, et al.. Towards Expert-Level Medical Question Answering with Large Language Models;. Available from: `http://arxiv.org/abs/2305.09617`.

[2] Meta Llama 3;. Available from: `https://llama.meta.com/llama3/`.

[3] Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, et al.. Llama 2: Open Foundation and Fine-Tuned Chat Models; 2023.

[4] Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, Salvi F, et al.. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models;. Available from: `http://arxiv.org/abs/2311.16079`.

[5] Axolotl - LLM fine tuning made easy;. Available from: `https://www.axolotlai.cloud/`.

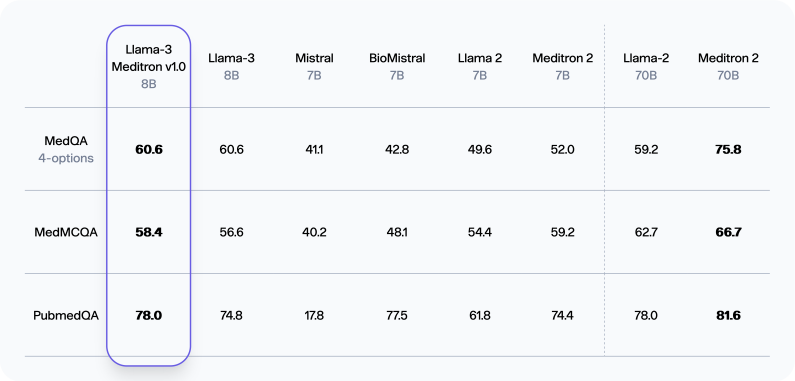
[6] Bosselut A, Chen Z, Romanou A, Bonnet A, Hernández-Cano A, Alkhamissi B, et al. MEDITRON: Open Medical Foundation Models Adapted for Clinical Practice;.

[7] Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics; 2019. p. 2567–2577. Available from: `https://aclanthology.org/D19-1259`.

[8] Pal A, Umapathi LK, Sankarasubbu M. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering; 2022.

[9] Labrak Y, Bazoge A, Morin E, Gourraud PA, Rouvier M, Dufour R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains;. Available from: http://arxiv.org/abs/2402.10373.

Appendix



	Llama-3 Meditron v1.0 8B	Llama-3 8B	Mistral 7B	BioMistral 7B	Llama 2 7B	Meditron 2 7B	Llama-2 70B	Meditron 2 70B
MedQA 4-options	**60.6**	60.6	41.1	42.8	49.6	52.0	59.2	**75.8**
MedMCQA	**58.4**	56.6	40.2	48.1	54.4	59.2	62.7	**66.7**
PubmedQA	**78.0**	74.8	17.8	77.5	61.8	74.4	78.0	**81.6**

Fig 3. Performance on MedQA, MedMCQA and PubmedQA across the various models