SOVEREIGN: How does the inference latency and memory footprint of MixLoRA compare to full fine-tuning on the MMLU benchma

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large language models (LLMs) can potentially democratize access to medical knowledge. While many efforts have been made to harness and improve LLMs' medical knowledge and reasoning capacities, the resulting models are either closed-source (e.g., PaLM, GPT-4) or limited in scale (<= 13B parameters), which restricts their abilities. In this work, we improve access to large-scale medical LLMs by releasing MEDITRON: a suite of open-source LLMs with 7B and 70B parameters adapted to the medical domain. MEDITRON builds on Llama-2 (through our adaptation of Nvidia's Megatron-LM distributed trainer)

1 Introduction

Analysis of: MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Research goal: How does the inference latency and memory footprint of MixLoRA compare to full fine-tuning on the MMLU benchmark for LLaMA-2 models at 7B, 13B, and 70B scales?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

7 papers retrieved. 9 claims extracted, 9 verified. Tribunal: 9.0/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
MEDITRON is a suite of open-source LLMs with 7B and 70B parameters adapted to the medical domain	✓	0.33
MEDITRON builds on Llama-2 through adaptation of Nvidia's Megatron-LM distributed trainer	✓	0.25
MEDITRON extends pretraining on a comprehensively curated medical corpus including selected PubMed articles, abstracts,	✓	0.35
Evaluations using four major medical benchmarks show significant performance gains over several state-of-the-art baselin	✓	0.29
MEDITRON achieves a 6% absolute performance gain over the best public baseline in its parameter class	✓	0.25
MEDITRON achieves a 3% absolute performance gain over the strongest baseline finetuned from Llama-2	✓	0.23
MEDITRON-70B outperforms GPT-3.5 and Med-PaLM compared to closed-source LLMs	✓	0.31
MEDITRON-70B is within 5% of GPT-4 compared to closed-source LLMs	✓	0.22
MEDITRON-70B is within 10% of Med-PaLM-2 compared to closed-source LLMs	✓	0.27

References

- https://doi.org/10.18653/v1/2025.findings-acl.68
- https://doi.org/10.48550/arxiv.2311.16079
- https://doi.org/10.1007/s11704-024-40663-9