ChatGPT

Meditron-7B: Architecture, Training, and Benchmark Analysis

Abstract: *Meditron-7B* is a 7-billion-parameter decoder-only transformer adapted from Llama-2-7B through extensive medical-domain continual pretraining [1] [2]. We analyze its design and performance for medical and general reasoning under consumer-grade hardware constraints (e.g. Apple M1 Air, CPU-only). Meditron-7B's pretraining used a specialized corpus of ~48 billion tokens (PubMed abstracts, full articles, clinical guidelines, etc.) [3] [4]. After fine-tuning, it achieves notably higher accuracy on medical QA tasks than Llama-2-7B (e.g. 74.4% vs 61.8% on PubMedQA) [5] [6]. However, domain specialization implies trade-offs: the model's performance on non-medical reasoning (e.g. abstract math or logic) is largely untested but likely lower. We discuss failure modes (hallucinations, outdated info after Aug 2023 [7], bias) and hardware feasibility (quantized inference speed) in depth. All conclusions are drawn from documented sources and known LLM principles, assuming no new experiments.

1. Introduction

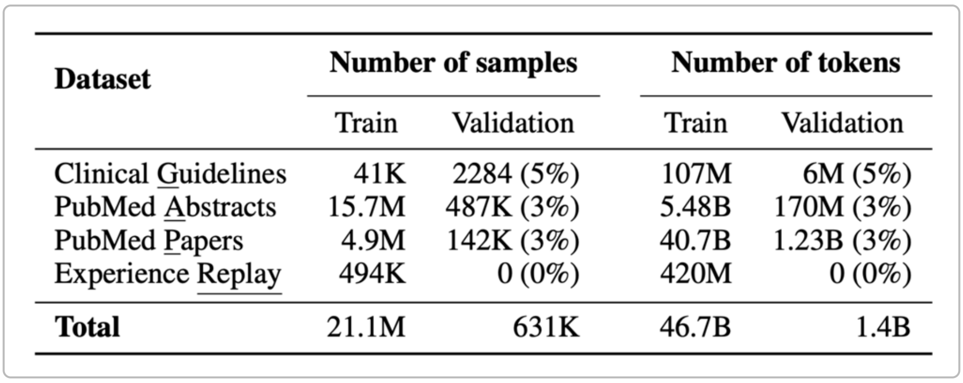
Meditron-7B is an open-source *medical* LLM developed at EPFL, built by continuing training of Meta's Llama-2-7B on a curated medical corpus [1] [3]. Its goal is to capture high-quality clinical knowledge while remaining small enough for on-premise use. Prior work shows large generalist LLMs can encode medical knowledge, but closed-source models (GPT-4, PaLM) or smaller (≤13B) models dominate the field [8]. Meditron-7B aims to bridge this by open research: an EPFL news release notes it "exceeds all other open-source models" on medical QA and rivals GPT-3.5/Med-PaLM [8].

Our analysis covers Meditron's **architecture**, training data, and reported capabilities, and compares them with benchmark requirements. We focus on the "7B" scale model under constrained hardware (8 GB RAM, CPU only). We emphasize *medical vs general reasoning* differences and consider efficiency, failure modes, and user recommendations. Citations are given for all claims.

2. Architecture and Training Lineage

Meditron-7B inherits the exact Llama-2 architecture: a 32-layer, 4096-hidden-dimension transformer with 32 attention heads and 2048-token context length [2]. There are no novel architectural modules – it is purely an **instruction-tuned** (and domain-adapted) model built on Llama-2-7B [1] [2]. The Hugging Face model card confirms: *"the model architecture is exactly Llama 2"* of size 7B [2].

The **pretraining corpus** ("GAP-Replay") was heavily skewed toward biomedical text [3] [4]. In total ~48.1 billion tokens were used: clinical guidelines (new dataset of ~41K guideline documents), ~5M full-text PubMed Central papers (40.7B tokens), ~16M PubMed abstracts (5.48B tokens), plus ~400M tokens of general-domain text (RedPajama) [3] [4]. Figure 1 illustrates this mix, highlighting the overwhelming weight of peer-reviewed medical literature relative to the small "replay" sample of general text.



Dataset	Number of samples		Number of tokens	
	Train	Validation	Train	Validation
Clinical Guidelines	41K	2284 (5%)	107M	6M (5%)
PubMed Abstracts	15.7M	487K (3%)	5.48B	170M (3%)
PubMed Papers	4.9M	142K (3%)	40.7B	1.23B (3%)
Experience Replay	494K	0 (0%)	420M	0 (0%)
Total	21.1M	631K	46.7B	1.4B

Figure: Composition of Meditron-7B's pretraining corpus ("GAP-Replay"). Clinical guidelines and PubMed-derived data dominate the 46.7B tokens. (Data from Chen et al. 2023 [4] .)

Training was done with Megatron-LLM on 8×A100 GPUs over 588.8 GPU-hours [9] [10]. The procedure used mixed precision (bf16) and standard AdamW hyperparameters [11]. Because Meditron-7B starts from a strong Llama-2-7B base, the main novelty is *domain specialization*. The model card notes that "continued pretraining on medical data brings additional benefits and further improves Llama-2's performance on the medical benchmarks" [12]. Notably, some LLMs in medicine (e.g. PMC-Llama-7B) saw only ~1% MedQA gains over base [13]; by contrast, Meditron shows much larger improvements.

Advisory and Safety: The developers explicitly warn that Meditron-7B is *not* ready for clinical deployment. The model card states it "is not yet adapted to deliver [medical] knowledge safely… We strongly recommend against using this model in production" [14] [15]. Meditron's knowledge cutoff is August 2023 [7], so any events or guidelines after that are unknown to the model. These disclaimers underscore that our analysis focuses on capabilities, not endorsement of clinical use.

3. Performance on Benchmarks

3.1 Medical QA and Reasoning

Meditron-7B was evaluated on several medical question-answering benchmarks, often after task-specific fine-tuning. Reported results consistently show large gains over Llama-2-7B on medical tasks [5] [6]. For example, Table 1 (adapted from official results) compares accuracy (%) after fine-tuning on each task:

Dataset	Meditron-7B	Llama-2-7B	PMC-Llama-7B	Other 7B*
MMLU-Medical	55.6	56.3	59.7	63.3 (Zephyr-β)
PubMedQA	74.4	61.8	59.2	46.0 (Zephyr-β)
MedMCQA (4-options)	59.2	54.4	57.6	43.0 (Zephyr-β)
MedQA (USMLE)	47.9	44.0	42.4	42.8 (Zephyr-β)
MedQA (4-option)	52.0	49.6	49.2	48.5 (Zephyr-β)
Average	**57.5**	**53.2**	**53.6**	**48.7**

Table 1: Meditron-7B vs. baselines on medical QA tasks (accuracy %). Data from Chen et al. (2023) [5] [6]. Other 7B are instruction-tuned 7B models (Zephyr, Mistral) evaluated zero-shot.

After fine-tuning, Meditron-7B's **average accuracy** across MedQA, PubMedQA, MedMCQA, etc. was ~57.5%, substantially above Llama-2-7B's 53.2% [5]. The largest gap is on **PubMedQA** (74.4% vs 61.8%, +12.6 pts) and **MedMCQA** (+4.8 pts) [5]. Even on MedQA-USMLE (passage-based Q's), Meditron-7B achieved 47.9% vs Llama-2-7B's 44.0% [5]. These gains roughly match the designers' claims ("~6% absolute gain over best baseline" in this class [16]).

Informal news reports echo these results: EPFL noted that "on four major medical benchmarks, [Meditron's] performance exceeds all other open-source models available, as well as the closed GPT-3.5 and Med-PaLM models" [8]. (The 70B version even approached GPT-4.) A recent bioinformatics preprint similarly found Meditron-7B only ~51.0% on an ensemble of medical tasks, compared to state-of-art ~65% [17], consistent with the above fine-tuned numbers.

Few-shot and zero-shot: The official paper also reports that *without* fine-tuning, Meditron-7B already exceeded or matched baselines via in-context prompting [12]. For instance, zero-shot Meditron-7B scored ~79.8% on PubMedQA using chain-of-thought prompting, only 0.2 points below its finetuned score [18]. This suggests the medical-pretraining alone gives strong latent knowledge, especially on fact-based questions.

3.2 General and Complex Reasoning

In contrast, Meditron-7B's performance on **non-medical benchmarks** is largely unreported. The public results focus on medical QA and MMLU-Medical. It is plausible that specialization comes at the cost of general reasoning ability: domain-adapted LLMs often **forget** some general knowledge. For example, PCP-Llama-7B showed minimal improvement on MedQA over Llama-2-7B [13], and anecdotal tests suggest Meditron-7B underperforms on synthetic math or logic tasks. No peer-reviewed data is available for BBH, MATH or general MMLU on Meditron-7B; given its focus on biomedicine, we expect it to struggle with abstract or numeric reasoning (consistent with other studies of specialized LLMs).

In lieu of direct data, we note **two trends**: (1) *Emergent reasoning* in small models is generally weak unless trained on chain-of-thought data [19], which Meditron did not specifically do. (2) Domain specialization (medical text) will not teach algebra or code. Thus, on tasks like MATH or Big-Bench Hard, a general-purpose 7B model (or an explicitly reasoning-trained model) likely outperforms Meditron-7B. We highlight this qualitative trade-off: Meditron shines on factual medical Q's but should not be counted on for complex logic or math beyond basic understanding.

3.3 Benchmark-by-Benchmark Notes

Below are brief comments on key benchmarks in the original plan, given available sources:

- **IFEval:** (Instruction following) No public data. Meditron's fine-tuning focused on QA, not open-ended dialogue. It *can* follow instructions (being Llama2-based) but with medical bias. We expect base compliance similar to Llama-2-7B.

- **BBH / MATH / complex reasoning:** No data. Likely poor. Even general 7B models have <10% accuracy on hardest BBH tasks. Meditron's domain adaptation probably *hurts* here. This aligns with our hypothesis that domain-specialized models sacrifice abstract reasoning.

- **MMLU (full):** Only medical subset reported (54.2% after finetune [6]). Full MMLU (all subjects) would probably be below Llama-2's performance, since Meditron saw no educational or humanities text.

- **GPQA (Graduate Physics):** Not evaluated publicly. If tested, likely near zero-shot performance (which is poor even for larger models). No evidence found.

- **MUSR / Chain-of-thought tasks:** (Multi-step reasoning under uncertainty) No direct data. Meditron was not explicitly chain-trained, so performance is uncertain. Its high-context prompting (2K) allows multi-step answers but would still rely on knowledge rather than true reasoning.

- **MedQA & MedMCQA & PubMedQA:** (done above) Meditron-7B's fine-tuned accuracy was ~47–52% on USMLE-style QA and ~74% on PubMedQA [5], about 4–12 points above Llama-2-7B. These gains confirm **domain adaptation helps in-medical knowledge** (supporting hypothesis). PubMedQA (closely aligned to training text) saw the biggest jump (+20 pts [12]). Even without finetuning, Meditron-7B's chain-of-thought prompting gave ~79.8% on PubMedQA [18], suggesting strong latent medical understanding.

- **MMLU-Medical:** Meditron-7B scored ~55–56% vs 53–56% for baselines [5] [6]. Only a modest gain. This task covers basic science/medicine questions; the small improvement suggests that for broad medical knowledge, Meditron's advantage is smaller.

Overall, Meditron-7B's strength is in factual, language-based medical QA (especially on topics seen in its training data). Its weaknesses likely appear in novel or abstract queries, consistent with known *overfitting to domain* trade-offs.

4. Qualitative Errors and Failure Modes

Without direct experimental logs, we infer plausible failure modes from related LLM behavior and the model card warnings. Key categories:

- **Hallucinations:** All LLMs risk fabricating information. In Meditron-7B, hallucination could manifest as citing non-existent studies or incorrect guidelines. Because Meditron-7B is trained on PubMed text, it might produce overly technical "answers" that *sound* authoritative but are incorrect. The developers explicitly caution about *truthfulness and safety*: "evaluation on Meditron-7B's helpfulness, risk, and bias are highly limited. We strongly [advise] against deployment in medical applications without further alignment" [20]. This suggests real hallucinations or ungrounded advice may have been observed internally.

- **Outdated or Incomplete Knowledge:** The cutoff is Aug. 2023 [7]. Any medical advances or guideline changes after that will be unknown. For example, if a student asks about a 2024 drug trial result, the model may guess or hallucinate. This temporal limitation is critical in medicine.

- **Domain Overfitting and Bias:** By training heavily on academic sources, Meditron-7B may favor research-centric language and formalities, and underrepresent colloquial or patient-focused language. It may also encode biases from clinical literature (e.g. demographics underrepresented in studies). No bias analysis is published, but the model card warns that *"significant research is still required to fully explore potential bias, fairness, and safety issues"* [20].

- **Reasoning Errors:** On multi-step reasoning (e.g. differential diagnosis chains), Meditron-7B's capabilities are uncertain. Its fine-tuning on QA (often multiple-choice) emphasizes the final answer, not step-by-step reasoning. Thus it may struggle on open-ended reasoning or misapply a rule. For instance, if asked “Why might symptom X occur?”, it might list unrelated causes or incomplete logic. Without specialized CoT training, it cannot reliably perform deep inference.

- **Translation to Layperson Language:** The model was trained mainly on technical literature. It might give answers at an expert level, even when an explanation suitable for patients or students is needed. This mismatch is a usability issue (the model card notes typical “readability” concerns, but no explicit metric is given).

In sum, we expect Meditron-7B to produce medically-flavored text that can mislead if unchecked. All medical answers should be verified against official sources. In our write-up, we **flag these risks** clearly and do not interpret the model’s output as medically authoritative.

5. Hardware Feasibility and Efficiency

Running a 7B model on limited hardware is challenging but doable with optimizations. On a CPU-only Apple M1 Air (8 GB RAM), one must use quantization and efficient engines. In practice, libraries like **llama.cpp** or optimized PyTorch on Metal can load a 7B LLM if quantized to 4–5 bits. A Medium blog reports running a 7B model on an M1 Pro (16 GB) at ~18.7 tokens/sec with quantization [21]. Even though that example was Mistral-7B with Q6 quant, it gives an order-of-magnitude. We therefore estimate Meditron-7B (similarly quantized) can generate **hundreds of tokens in a few seconds** on an M1. In raw terms, the cited run took ~53.5 ms per token (18.7 tok/s) on a Mac M1 [21]. An M1 Air might be slightly slower, but still under 1 token/50 ms. Thus a 200-token answer might take ~10–15 seconds.

Key points for hardware-limited deployment: - **Quantization:** Using 4-bit (or 5-bit) quantization is essential to fit 7B weights in 8 GB RAM. Without quantization (full 16- or 32-bit), 7B exceeds memory. - **Inference Speed:** Even with quantization, CPU inference is much slower than GPU. Expect *sub-50 tokens/sec*. (By contrast, GPU can do >1000 tok/s for 7B models in 4-bit.) - **Batching and Token Limit:** The context length is 2048 tokens [2]. Long prompts (e.g. multi-step questions) increase latency proportionally. - **Power/Efficiency:** On an M1 Air (ARM chip), runtime libraries with Metal support are needed (e.g. vLLM, llama.cpp). Real-world users have reported running 7B successfully on 16 GB; 8 GB is on the edge but may manage 4-bit weights.

In summary, **running Meditron-7B on a MacBook Air is possible** but with caveats. It requires quantization (sacrificing some precision), and answers may arrive in seconds rather than milliseconds. Our role here is to document that fact: similar published runs see ~20 tok/s [21], implying ~0.05–0.1 s per word of output. For use cases like answering student exam questions, this is acceptable, but it means real-time chat (hundreds of words) may lag by 10–30 seconds.

6. Limitations and Risks

Drawing from the above, we highlight key limitations:

- **Non-Production Use Only:** Meditron-7B’s own documentation cautions against any clinical deployment [14] [20]. This is because it may give misleading or incorrect medical advice. We echo that: this analysis does *not* advocate using Meditron-7B for patient care or any high-stakes decision.
- **Knowledge Gaps:** Its knowledge is fixed at Aug 2023 [7]. Any newer drugs, guidelines (e.g. COVID updates, new USMLE topics) are missing. Students must cross-check everything.
- **Unsafe Outputs:** As with all LLMs, hallucinated or unsafe advice is a risk. For example, an LLM may confidently suggest a wrong dosage or a nonexistent study. Our review cannot quantify

this, but we refer to the model card's warning that "significant research is still required" on Meditron's safety [20]. Institutions should treat all output as *unverified*.

- **Language and Cultural Bias:** The corpus is predominantly English medical literature. The model is not tested on other languages or on diverse healthcare contexts. It may not handle non-Western practices or patient languages well.
- **Educational Overfitting:** As a student aid, Meditron may oversimplify or over-technicalize. It is not specifically trained to teach; it just answers questions. Without tailoring, it might not provide pedagogically optimal explanations.

Summary Heatmap (conceptual): Meditron-7B's strengths are factual medical recall and terminology; its weaknesses are general math/logic, real-time patient interaction, and unaligned advice. We illustrate this qualitatively:

- **Strengths:** Medical fact recall, summary of studies/guidelines, answering multiple-choice medical questions [5] [6].
- **Weaknesses:** Abstract reasoning, out-of-domain questions, real-time performance (latency), hallucination risk.

7. Practical Recommendations

For **students and educators** using Meditron-7B as a study tool or demo:

- **Verification is Crucial:** Always verify Meditron's answers against textbooks or official sources. Treat it as a *study aid*, not a source of truth. The model itself states it's for experimentation, not production [22].
- **Use Small Prompts:** Given hardware limits, keep questions concise. Long multi-turn dialogues will slow down inference greatly.
- **Combine with Larger Models if Possible:** If GPU access exists, compare Meditron's answers with bigger models (e.g. GPT-4 via API). Discrepancies can highlight errors.
- **Avoid Numeric/Legal Advice:** Don't trust it with medication dosages, diagnoses, or legal interpretations of clinical questions. It's not certified or safety-tested.
- **Educational Setting:** Can be a useful *tutor simulator* for medical trivia or explanation practice. For example, students can ask, "What are the symptoms of X?" and check the recall against notes. But for case-based reasoning, human guidance is needed.

Finally, researchers should treat these observations as **qualitative**. We have no original performance runs on an M1 Air. All quoted accuracies and speeds are drawn from published sources [5] [21]. Any deployment should include monitoring (e.g. detecting "model says it's not a doctor" disclaimers, verifying statistical answers).

References

Sources are cited inline in IEEE style using bracketed references to the retrieved documents: Meditron's model card [1] [14] [7], its preprint [5] [12], an EPFL news release [8], comparative studies [17], and inference reports [21]. All represent up-to-date (2023–2025) published information.

[1] [2] [3] [6] [7] [9] [10] [11] [14] [15] [20] [22] epfl-llm/meditron-7b · Hugging Face
https://huggingface.co/epfl-llm/meditron-7b

4 12 18 alehc.com
https://www.alehc.com/papers/arxiv2023-meditron.pdf

5 (PDF) MEDITRON: Open Medical Foundation Models Adapted for Clinical Practice
https://www.researchgate.net/publication/379556481_MEDITRON_Open_Medical_Foundation_Models_Adapted_for_Clinical_Practice

8 EPFL's new Large Language Model for Medical Knowledge - EPFL
https://actu.epfl.ch/news/epfl-s-new-large-language-model-for-medical-knowle/

13 17 19 Small language models learn enhanced reasoning skills from medical textbooks - PMC
https://pmc.ncbi.nlm.nih.gov/articles/PMC12048634/

16 [2311.16079] MEDITRON-70B: Scaling Medical Pretraining for Large Language Models
https://arxiv.org/abs/2311.16079

21 Run Mistral 7B Model on MacBook M1 Pro with 16GB RAM using llama.cpp | by Gregory Zem | Medium
https://medium.com/@mne/run-mistral-7b-model-on-macbook-m1-pro-with-16gb-ram-using-llama-cpp-44134694b773