sustainability

MDPI

Article

MediLITE-QA: Energy-Efficient Medical Question Answering with a Quantized Domain-Specific Language Model on a Commodity Laptop CPU

Olusegun Oladele Odewole [1,*] and Seren Başaran [2]

[1] Department of Computer Information Systems, Faculty of Artificial Intelligence and Informatics, Near East University, Mersin 10, Lefkoşa 99138, Turkey
[2] Department of Data Analytics Engineering, Faculty of Artificial Intelligence and Informatics, Near East University, Mersin 10, Lefkoşa 99138, Turkey; seren.basaran@neu.edu.tr
* Correspondence: 20225054@std.neu.edu.tr

Abstract

Frontier large language models (LLMs) depend on energy-intensive cloud data centers and continuous connectivity that many low-resource healthcare settings lack. This paper investigates whether Low-Rank Adaptation (LoRA) fine-tuning combined with 4-bit GGUF quantization enables accurate medical question answering (QA) on commodity CPUs without GPU acceleration or cloud connectivity. We instantiate this pipeline as MediLITE-QA, a fine-tuned Phi-3.5-mini-instruct (3.8B-parameter) model. On a consumer-laptop CPU, the deployed model reaches a steady-state Time to First Token of 1.34 s, 8.19 tokens/second, and 71.87% average accuracy across nine medical QA benchmarks (72.79% before quantization, not significantly different from the 8B model OpenBioLLM-8B). The 7–8B comparison models used 1.56–1.98× more CPU energy per token and showed cold-start penalties that MediLITE-QA did not. Fine-tuning GPU-board energy, estimated from power telemetry, was approximately 139 kWh. In an exploratory blinded clinical evaluation, the fine-tuned model was not significantly better than its base model and remained significantly below a frontier cloud reference (Wilcoxon $p_{adj} = 0.0023$, $n = 100$). On this hardware tier, the pipeline is a viable, lower-energy-per-token alternative to larger models, relevant to Sustainable Development Goals 3, 9, and 10; results should not be assumed to transfer to phones, single-board computers, or servers, and accuracy gaps must close before clinical use.

Keywords: sustainable artificial intelligence; green computing; CPU-only inference; energy-efficient inference; on-device large language models; model quantization; LoRA fine-tuning; digital health equity; healthcare accessibility; resource-constrained systems

Check for updates

Academic Editor: Andreas Kanavos

Received: 27 September 2026
Revised: 22 October 2026
Accepted: 26 October 2026
Published: 4 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sustainable digital transformation of healthcare requires artificial intelligence (AI) systems that are not only accurate but also resource-efficient, energy-conscious, and deployable without dependence on centralized cloud infrastructure that many health systems—and most of the world's population outside high-income countries—cannot reliably access [1]. Large language models (LLMs) are increasingly proposed as clinical decision-support and patient-facing tools, but the dominant deployment paradigm couples them to GPU-backed cloud data centers whose energy consumption, cooling-water use, and embodied carbon footprint are substantial and growing, and whose continuous-connectivity requirement is itself a sustainability and equity liability in bandwidth-constrained or intermittently

connected settings [2]. This tension between AI capability and environmental and infrastructural sustainability is a defining challenge for responsible AI deployment in healthcare, and motivates a research question with direct relevance to Sustainable Development Goal (SDG) 3 (Good Health and Well-Being), SDG 9 (Industry, Innovation and Infrastructure), and SDG 10 (Reduced Inequalities): can domain-specific medical LLMs be made accurate enough to be clinically useful while running entirely on low-power, commodity CPU hardware, eliminating both the energy overhead of dedicated AI accelerators and the connectivity dependency of cloud inference?

This question is driven by three converging sustainability pressures. First, *energy and infrastructure efficiency*: GPU-accelerated cloud inference concentrates energy demand in data centers and requires capital- and energy-intensive accelerator hardware that is economically and environmentally costly to deploy at the scale needed for broad healthcare access, particularly in resource-constrained health systems where such infrastructure investment competes directly with core clinical spending. CPU-only, quantized on-device inference instead reuses existing, already-amortized commodity computing hardware, avoiding additional accelerator hardware [3]; whether it also lowers operational energy relative to cloud inference depends on factors this study does not measure (Section 4.4). Second, *equitable access and the digital divide*: reliable, high-bandwidth connectivity to cloud AI services cannot be assumed in rural clinics, low- and middle-income-country (LMIC) health systems, or disaster and humanitarian response settings, precisely the contexts where AI-assisted triage and information support could have the greatest marginal benefit; offline, on-device inference removes this dependency and is a sustainability-of-access requirement rather than a convenience. Third, *data sovereignty and long-term system resilience*: transmitting patient-related queries to third-party cloud LLM services is frequently undesirable or non-compliant with data protection requirements, and health systems that build clinical workflows around a single external cloud provider create a fragile, externally dependent architecture that is itself a long-term sustainability risk for critical health infrastructure. Recent advances in parameter-efficient fine-tuning (e.g., Low-Rank Adaptation, LoRA [4]) and post-training quantization (e.g., GGUF-format weight quantization as implemented in inference engines such as llama.cpp [5]) have narrowed the technical gap that previously made local, CPU-only LLM deployment impractical, but systematic, reproducible characterizations of the resulting efficiency-accuracy trade-offs—and of what these trade-offs mean for sustainable, equitable healthcare AI deployment—remain limited [3].

This paper uses medical question answering (QA) as a case study domain to develop and evaluate such a pipeline, for reasons specific to sustainable healthcare AI. First, healthcare is a domain where both the environmental-efficiency and the access-equity motivations for on-device inference are unusually concrete: every additional layer of dedicated accelerator hardware and every additional dependency on continuous cloud connectivity is a barrier to deploying AI-assisted clinical support in precisely the lower-resource settings that stand to benefit most, while also adding to the aggregate energy footprint of the health sector's digital infrastructure. Second, medical QA carries meaningfully high accuracy and safety stakes, providing a demanding testbed for evaluating whether an efficiency-optimized, low-resource deployment pipeline retains acceptable task performance—a question directly relevant to whether resource-light AI can be a genuine sustainability solution in healthcare, or only a lower-quality substitute that trades accuracy for efficiency in a way that is not clinically acceptable. We emphasize that this paper's central contribution is the deployment pipeline and its systematic efficiency-accuracy characterization on CPU-only hardware, together with an explicit assessment of its sustainability and equity implications; the resulting model, MediLITE-QA, is presented as a concrete instantiation of that pipeline rather than

as a clinically validated medical product (see Section 4.5 for an explicit discussion of this distinction and the further validation required before any clinical application).

General-purpose LLMs such as GPT-4 have significantly advanced natural language processing by training on large-scale text corpora. However, in specialized domains like healthcare, the strongest general-purpose models, such as ChatGPT, achieve high benchmark accuracy only at a scale that requires cloud hosting, and are not adapted to specific clinical settings [6]. Because they are proprietary and cloud-hosted, they also offer limited transparency and reproducibility, are harder for lower-resource health systems to access, and leave their sustainability profile entirely dependent on how external providers choose to operate them [7].

One response to these challenges has been domain-specific language models, fine-tuned on specialized datasets to better capture medical language and reasoning. Wu et al. [6] continually pretrained a LLaMA-based model on biomedical literature and obtained substantially higher medical QA accuracy than its general-purpose base model. Kim et al. [8] reported comparable reasoning gains in smaller models fine-tuned directly on medical textbooks instead of general web text, and Zhang et al. [9] found that domain fine-tuning can outperform general-purpose models of similar size on medical consultation benchmarks.

Because transformer language models are architecturally complex and have large parameter counts, both training and inference require substantial computation and therefore substantial energy [10,11]. Smaller and more efficient models now exist, but Rohanian et al. [12] found that compact biomedical transformers still trail larger models on complex reasoning tasks and remain limited in contextual understanding and adaptability on resource-constrained devices. This work addresses these gaps with a compact, quantized model for medical QA that retains competitive benchmark accuracy and can run in low-resource, low-energy settings.

ClinicalXLNet [13] and BioALBERT [14] target tasks such as disease prediction and named entity recognition; both are primarily encoder-based and are not suited to generative QA. Luo et al. [15] introduced BioGPT, a GPT-2-scale model pretrained on biomedical abstracts, which must be fully retrained to adapt to a new subdomain. Kanakarajan et al. [16] proposed BioELECTRA, whose discriminator-based pretraining improves sample efficiency, although the model still targets encoder-style tasks rather than open-ended generation. Lu et al. [17] developed ClinicalT5, which supports generation, but its computational cost, like that of the other two, limits real-world deployment on edge devices. Systems like ChatDoctor [18], HuatuoGPT [9], Zhongjing [19], GatorTronGPT [20], and BioMistral [21] provide improved performance in medical QA, but with large parameter sizes (7B–70B) they carry a larger inference-time resource footprint: even where the smaller of them run on commodity hardware, they leave little memory headroom and cost more energy per generated token, and the largest exceed it. Few models are simultaneously generative, domain-specific, efficient enough for CPU-based or mobile/edge deployment, and evaluated with explicit attention to both resource footprint and accuracy.

Table 1 summarizes the scale and architecture of a representative set of open, generative medical LLMs. What is consistently absent across this prior work, regardless of model size, is a systematic CPU-only efficiency characterization (latency, throughput, memory, and energy) comparable to that reported in Sections 3.1 and 3.2. GGUF conversion itself is not a meaningful differentiator, as it is a mechanical, publicly documented post-training step that can be applied by any practitioner to most open-weight models regardless of whether the original authors released one; what is actually missing from prior work is the systematic resource-efficiency characterization once that conversion is done, and its interpretation through a sustainability and equitable-access lens rather than accuracy alone.

Table 1. Base architecture and scale for representative open generative medical LLMs. The proposed model, MediLITE-QA, is shown in bold.

Model	Base Architecture	Params
PMC-LLaMA [6]	LLaMA-1	7B/13B
ChatDoctor [18]	LLaMA	7B
HuatuoGPT [9]	BLOOMZ-7B1-mt	7B
BioMistral [21]	Mistral-7B	7B
OpenBioLLM-8B [22]	Llama-3-8B	8B
MediLITE-QA (ours)	Phi-3.5-mini-instruct	3.8B

Beyond weight-based parameter-efficient adaptation such as LoRA, learnable prompt representations offer an alternative mechanism for domain adaptation: prompt tuning [23] and prefix tuning [24] keep the base weights frozen and learn only a small set of continuous prompt vectors. In medicine, Cai et al. [25] proposed MedLP-HAFB-CLIP, a hierarchical adaptive framework in which learnable medical prompts, enriched with fine-grained anatomical descriptions, adapt a pretrained vision–language model to fetal ultrasound standard-plane identification. Although its task and modality differ from text-based medical question answering, its underlying principle—introducing domain knowledge through trainable prompt representations rather than weight updates alone—could support specialty- or language-specific adaptation of compact medical language models with few trainable parameters. Its deployment implications differ from those of a merged LoRA adapter, however, because soft prompts lengthen every input and may therefore raise time to first token and per-query energy on CPU-only systems; Section 4.7 discusses this complementary trade-off in detail.

Despite this substantial body of work, four specific gaps remain. (1) Prior domain-specific medical LLMs are evaluated primarily on accuracy benchmarks, with limited systematic reporting of the latency, throughput, memory, and energy footprint trade-offs incurred when the same model is deployed under genuine CPU-only constraints. (2) Comparisons between fine-tuned domain models and both their own unquantized base models and frontier cloud references are frequently reported using inconsistent evaluation protocols, making it difficult to isolate the effect of fine-tuning and quantization from the effect of prompting differences. (3) Existing edge-LLM deployment studies rarely distinguish explicitly between genuinely mobile/embedded hardware and higher-resource CPU-only server infrastructure, despite the two presenting substantially different resource and energy profiles. (4) Prior work rarely connects efficiency-oriented LLM deployment choices to their implications for sustainable and equitable healthcare access, treating computational efficiency as an engineering property rather than a determinant of who can practically deploy and benefit from the technology. Accordingly, this study addresses three research questions:

- **RQ1.** To what extent does domain-specific QLoRA fine-tuning improve medical QA accuracy over the Phi-3.5-mini base model, and how much of this gain is retained after Q4_K_M quantization?
- **RQ2.** What latency, throughput, memory, and CPU-energy trade-offs does the deployed model exhibit relative to directly evaluated larger open models on a resource-constrained consumer-laptop CPU?
- **RQ3.** What clinical-correctness and safety limitations emerge in an exploratory comparison with the base model and a frontier cloud reference, and what do they imply for the model's current sustainability and clinical-readiness claims?

This paper's contributions, addressing these four gaps and three research questions, are fivefold:

- An efficient, reproducible pipeline combining LoRA fine-tuning and 4-bit GGUF quantization to enable interactive-latency, GPU-free inference on a consumer-laptop CPU for a domain-specific medical LLM;
- A systematic characterization of the resulting model against same-class and larger alternatives—steady-state and cold-start latency, throughput, memory footprint, and directly measured CPU energy on identical CPU-only hardware, together with protocol-matched benchmark accuracy;
- An explicit accounting of what this evaluation does and does not establish about mobile/embedded (as opposed to consumer-laptop or server-class) deployment;
- A domain-specific case study—medical question answering—assessing the pipeline's output quality via both automated and exploratory blinded human clinical evaluation, illustrating a methodology applicable to other high-stakes, resource-constrained application domains;
- A discussion of the pipeline's implications for sustainable, low-carbon, and equitable healthcare AI deployment, framed against SDG 3, SDG 9, and SDG 10, including an explicit account of the accuracy and safety limitations that currently bound its readiness as a sustainability solution.

Scope Statement: Consumer-Laptop CPU vs. Mobile/Embedded Hardware

All latency, throughput, memory, and energy measurements reported in this paper were performed on a dedicated, non-virtualized consumer laptop (11th Gen Intel Core i5-1135G7, 4 physical cores/8 threads, 6.96 GiB RAM; see Section 2 for full hardware disclosure), not a cloud instance. This device is a genuinely resource-constrained deployment target in its own right: the 7–8B-parameter comparison models evaluated occupy 84–89% of its total RAM at peak (Sections 3.5 and 4.4), a condition we show has direct, measurable latency consequences. This is nonetheless a materially different resource profile from both (a) genuinely mobile/embedded hardware (smartphones, single-board computers) and (b) a high-memory, multi-core bare-metal server: the laptop's 4 physical cores and 6.96 GiB RAM place it above typical mobile/embedded specifications but below a dedicated high-memory server. We therefore characterize our evaluation as representative of a resource-constrained *consumer laptop* CPU tier, distinct from both. This paper does *not* report results on phones, tablets, wearables, single-board computers, or dedicated high-memory bare-metal servers, and the reported latency/throughput/memory figures should not be assumed to transfer directly to such devices, which face additional or different constraints (battery/power budget, different RAM configurations, virtualization overhead where applicable) not captured by our benchmark environment. Direct energy measurement for this configuration, using the processor's Intel RAPL counters, is reported in Section 3.5.

The remainder of this paper is organized as follows. Section 2 details the fine-tuning, quantization, and evaluation-suite construction methodology. Section 3 reports the resulting latency/throughput/accuracy characterization, the automated cross-model comparison, the blinded human clinical evaluation, and the qualitative error analysis. Section 4 discusses privacy implications, sustainability and equity implications, clinical-readiness limitations, and the human-evaluation design limitations. Section 5 concludes with the paper's practical takeaways and directions for future work.

2. Materials and Methods

Figure 1 provides a visual overview of the full deployment and evaluation pipeline described in this section: base model selection, LoRA fine-tuning, checkpoint merging, 4-bit GGUF quantization, CPU-only inference, and the parallel evaluation tracks (automated benchmark accuracy, automated cross-model metrics, blinded human clinical evaluation,

and latency/throughput/memory/energy characterization) whose results are reported in Section 3. Only the efficiency track runs on the CPU; accuracy and the cross-model and human evaluations were run on GPU (Sections 2.4 and 2.6).

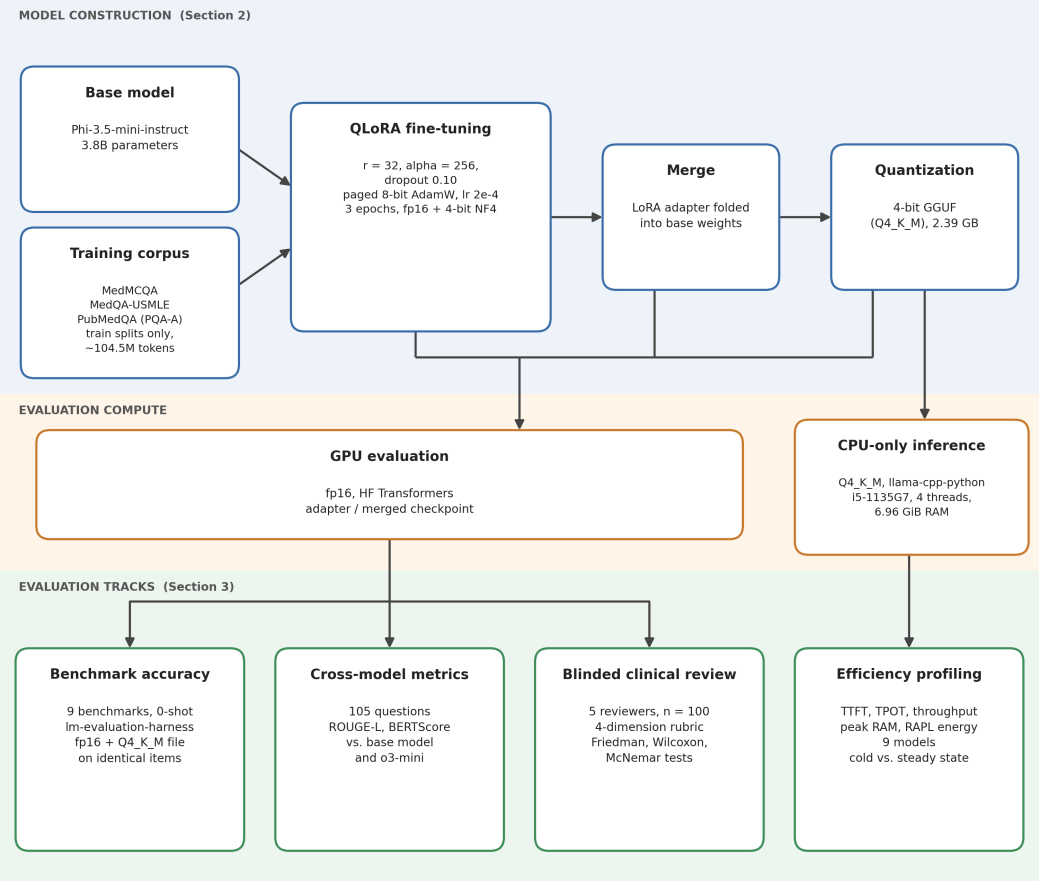


Figure 1. Overview of the MediLITE-QA pipeline: QLoRA fine-tuning of Phi-3.5-mini-instruct, merging, and 4-bit GGUF quantization, followed by two evaluation compute paths—GPU (fp16) for benchmark accuracy, cross-model metrics, and blinded clinical review, and CPU-only inference of the Q4_K_M file for efficiency profiling (Section 3).

The development of MediLITE-QA, a compact, domain-specific transformer for medical question answering, followed a four-stage process optimized for CPU-only deployment on consumer-laptop hardware. First, `microsoft/Phi-3.5-mini-instruct` (3.8B parameters) was selected for its efficient, open-source architecture and support for long-context processing at low computational demand. Second, a training corpus of approximately 104.5 million tokens (Phi-3 tokenizer; Table 2) was compiled exclusively from the official training splits of MedMCQA (`openlifescienceai/medmcqa`), MedQA-USMLE-4-options (`GBaker/MedQA-USMLE-4-options`), and PubMedQA (`bigbio/pubmed_qa`), covering diagnostic multiple-choice reasoning, USMLE-style clinical vignettes, and biomedical research-abstract reasoning respectively. Preprocessing formatted each training item into the Phi-3.5 chat template (question and lettered answer options as the user turn, the answer letter as the assistant turn) and tokenized it with the Phi-3 tokenizer; no lowercasing was applied. Third, the model was fine-tuned using Low-Rank Adaptation (LoRA) by updating low-rank matrices in the transformer's attention and feedforward layers, with the base model loaded in 4-bit NF4 (i.e., QLoRA). Training used the paged 8-bit AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, learning rate $= 2 \times 10^{-4}$) for 3 epochs with fp16 mixed precision on a single NVIDIA Quadro RTX 4000 GPU, which does not support bf16. Finally, accuracy was evaluated on nine medical benchmarks with EleutherAI's evaluation harness, and the deployed

GGUF Q4_K_M files were scored on the identical logged requests (Section 2.4). All latency, throughput, memory, and energy measurements were conducted on a CPU-only system using GGUF Q4_K_M quantized models via the llama-cpp-python library with GPU layers disabled. The benchmark host is a dedicated, non-virtualized consumer laptop (11th Gen Intel Core i5-1135G7, Tiger Lake, 4 physical cores/8 threads via 2-way hyperthreading, 6.96 GiB total RAM), whose scope is discussed in the Scope statement (Section 1); Section 3.2 reports the resulting on-disk model size and peak RAM usage alongside the latency figures in Figure 2, measured under this same configuration; the corresponding per-query energy consumption is reported in Section 3.5.

2.1. Dataset Statistics and Composition

To ensure reproducibility, we detail the composition of our 104.5-million-token training corpus in Table 2, with token counts measured using the Phi-3 tokenizer. Training used exclusively the official training splits of each dataset; the held-out MedQA and PubMedQA test sets and the MedMCQA validation set (whose test labels are not public; Section 2.4), together with the medical subsets of MMLU (Clinical Knowledge, Medical Genetics, Anatomy, Professional Medicine, College Biology, College Medicine), constitute the nine-benchmark evaluation suite reported throughout Section 3. The corpus is dominated by PubMedQA (86.1% of total tokens), reflecting that dataset's longer, abstract-length reasoning passages relative to the shorter multiple-choice item stems of MedMCQA and MedQA-USMLE. The 200,000 PubMedQA training records are the artificially labelled PQA-A subset, whose labels were generated automatically and take only yes/no values; the "maybe" class therefore never occurs in fine-tuning, which is relevant to the PubMedQA results in Sections 3.1 and 3.6.

Table 2. Fine-tuning corpus composition (Phi-3 tokenizer token counts). All 393,000 records come from the official training splits of the three source datasets. The bottom rows show the random 95%/5% split of this corpus into records used for gradient updates and records held out for validation loss during training; the benchmark evaluation sets are not part of this corpus.

Dataset	HuggingFace ID	Total Tokens	% of Corpus
PubMedQA	`bigbio/pubmed_qa`	89,910,106	86.06%
MedMCQA	`openlifescienceai/medmcqa`	11,967,710	11.46%
MedQA-USMLE-4-options	`GBaker/MedQA-USMLE-4-options`	2,596,338	2.49%
Total (all datasets)		104,474,154	100.00%
Training records, 95% (373,350)		99,240,768	94.99%
Validation records, 5% (19,650)		5,233,386	5.01%

2.2. Data Quality and Bias Limitations

We acknowledge that publicly available medical QA datasets, drawn from examination question banks and published abstracts, often contain demographic biases and potential label noise. As noted by Rajkomar et al. [26], such data may underrepresent non-Western populations or reflect outdated clinical practices, a concern directly relevant to the equitable-access aims of this work (Section 4.3): a model intended to widen access to medical QA in underserved settings must not simultaneously encode systematic bias against the populations it is meant to serve. No deduplication or other cleaning was applied beyond formatting and tokenization (Section 2); overlap between the training corpus and the evaluation sets was assessed post hoc (Section 2.4). We also did not perform specific demographic filtering in this iteration. Future work will incorporate diverse sourcing to mitigate these inherent biases.

2.3. Hyperparameters

Formally, LoRA [4] re-parameterizes the update to a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ during fine-tuning as a low-rank decomposition rather than a full-rank update:

$$h = W_0 x + \Delta W x = W_0 x + BAx, \quad (1)$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d,k)$ bounds the number of trainable parameters introduced per adapted layer to $r(d+k)$ rather than the dk parameters a full fine-tuning update would require, which reduces the trainable parameters and optimizer state relative to full fine-tuning; we did not measure full fine-tuning for an energy comparison. A is initialized from a random Gaussian distribution and B is initialized to zero, so that $\Delta W = BA = 0$ at the start of training and the adapted model is identical to the base model before any fine-tuning steps are taken.

We employed LoRA with a rank of $r = 32$, scaling factor $\alpha = 256$, and a dropout of 0.10. Training additionally used a weight decay of 0.1 and a 100-step linear warmup on the learning rate schedule (Table A1 in Appendix A consolidates the full configuration). These values were chosen heuristically, balancing trainable parameter count against model expressivity, and were not tuned or ablated in this study; a sensitivity sweep across LoRA ranks (e.g., 16 vs. 32 vs. 64), together with a corresponding energy-per-rank comparison, is left to future work.

2.4. Data Contamination and Evaluation Setup

To benchmark model performance, we used the EleutherAI `lm-evaluation-harness` framework, standardizing our evaluation of MediLITE-QA and directly-evaluated open-source comparison models under a strict 0-shot setting. This configuration was intentionally selected to test the models' inherent biomedical knowledge and reasoning capabilities, removing the variance introduced by in-context learning or few-shot exemplars. Each answer option is scored by its log-likelihood given the question, and the highest-scoring option is taken as the prediction; no text is generated, so sampling parameters do not affect accuracy. As in the standard harness protocol, no chat template is applied for any model, so that accuracy comparisons are not confounded by model-specific prompt formats; because MediLITE-QA was fine-tuned on chat-formatted data, this protocol may slightly understate its accuracy relative to a chat-formatted evaluation. The fine-tuning corpus draws exclusively from the train splits of MedMCQA, MedQA-USMLE-4-options, and PubMedQA (Section 2.1). Evaluation uses the MedQA test split ($n = 1273$), the PubMedQA labelled test split ($n = 500$), the MedMCQA validation split ($n = 4183$; MedMCQA test labels are not public), and six medical subsets of MMLU, which is not present in the training corpus at all. Table 3 lists the complete evaluation and inference configuration.

Contamination screen. Independently of the publishers' splits, we screened all 7045 evaluated items against the complete fine-tuning corpus (393,000 records, before the train/validation split) using word 13-gram matching, the length used in the GPT-3 contamination analysis [27], after Unicode normalization, lower-casing, and punctuation removal. Each evaluation item's full text (question, context, and answer options, exactly as scored) was decomposed into 13-grams, which were indexed and checked against every training record. Adapting the token-coverage approach used for Llama 2 [28], an item's coverage is the share of its words lying inside any 13-gram also present in the training corpus; an item was classed as *dirty* if coverage was ≥80% or its question stem was duplicated (every 13-gram of the stem, or every 8-gram for stems of 8–12 words, present in the corpus), and as *not clean* if coverage was ≥20% or the question stem was duplicated. Dirty items made up 0.57% of the evaluation set (MedMCQA 32, MedQA 6, MMLU Anatomy 1, MMLU Clinical Knowledge 1; PubMedQA 0). Partial overlap (≥20%) was concentrated

in USMLE-style vignettes (MedQA 19.2%, MMLU Professional Medicine 16.2%) with no duplicated question stems (per-benchmark rates in Supplementary Table S3); inspection of the most frequent matching 13-grams (Supplementary Table S3c) showed standard item templates—question stems ("which of the following is the most likely underlying cause of this patient's symptoms"), examination boilerplate ("the remainder of the examination shows no abnormalities"), and vital-sign strings—rather than item-specific clinical content, and its presence in MMLU Professional Medicine, which is not part of the training data, likewise indicates shared template phrasing rather than item leakage. Excluding dirty items left the fine-tuning effect unchanged (+4.05 pp, 95% CI 2.40–5.70), and excluding all not-clean items reduced it only slightly (+3.77 pp, 95% CI 2.08–5.46; Supplementary Table S3b); MedQA and MedMCQA improvements remained significant in both analyses. Because 13-gram matching detects only near-verbatim overlap, this screen is a necessary rather than a sufficient contamination check: paraphrased overlap between training and evaluation items cannot be excluded and remains a residual limitation (Section 4.5).

Table 3. Evaluation and inference configuration.

Setting	Value
Accuracy (all directly evaluated models)	
Framework	`lm-evaluation-harness` 0.4.11, `transformers` 4.56, `datasets` 3.2.0
Precision/hardware	fp16 on 2× NVIDIA T4 GPU; batch size 4
Protocol	0-shot; log-likelihood scoring of answer options; no chat template; metric `acc` (identical to `acc_norm` for the letter-choice tasks)
Deployed-model accuracy	Q4_K_M GGUF files scored on the identical logged (context, answer-option) requests with llama.cpp using GPU offload (deterministic scoring; the device affects only numerical rounding); scorer validated against an independent full-sequence evaluation (identical option choice on 40/40 validation items)
Latency, throughput, memory, and energy (CPU-only)	
Engine	llama-cpp-python (0.3.35 for the energy runs), GPU layers 0
Threads/context	4 threads (one per physical core); $n_{ctx} = 2048$
Decoding	Greedy (temperature 0); max. 256 new tokens; repeat penalty 1.0 (top-p 0.95, top-k 40, min-p 0.05 are library defaults, inactive under greedy decoding)
Prompt format	Each model's own chat template (MediLITE-QA and Phi-3.5: template embedded in the GGUF file; OpenBioLLM-8B: Llama-3; BioMistral-7B: Mistral-Instruct; Apollo-7B: Alpaca; AlpaCare: Llama-2; Gemma 2B: Gemma); LLaMA 3 8B and ClinicalGPT (base models): plain "Question: ... Detailed Medical Answer:" prompt

Statistical analysis. Per-benchmark accuracies are reported with Wilson 95% confidence intervals and nine-benchmark averages with item-level bootstrap intervals (10,000 resamples). Paired comparisons between models use exact McNemar tests on each benchmark's discordant items, Holm-corrected across the nine benchmarks within each comparison, and a paired item-level bootstrap for the difference in average accuracy.

2.5. Custom High-Risk Evaluation Suite

To support both the automated cross-model evaluation (Section 3.3) and the human expert evaluation (Section 3.4), we additionally authored a set of 25 clinical safety questions covering high-stakes scenarios in which an incorrect answer carries significant potential for patient harm: drug-drug and drug-condition interactions (e.g., anticoagulant/antiplatelet combinations, CYP450-mediated interactions), absolute contraindications (e.g., medica-

tions contraindicated in pregnancy, neonates, or renal impairment), and emergency presentations requiring time-critical recognition (e.g., sepsis, anaphylaxis, aortic dissection, hyperkalemia). Each item was authored with a corresponding reference answer specifying the clinically correct action and its rationale. This suite is distinct from, and not part of, the training data described in Section 2.1; it was constructed solely for held-out evaluation, to specifically probe safety- critical failure modes that standard multiple-choice medical QA benchmarks (MedMCQA, PubMedQA) do not directly target. The full suite, with reference answers, is provided in Supplementary Table S1.

2.6. Human Evaluation Design

Five clinical reviewers were recruited via direct invitation from the authors' professional network. The review panel consisted of five general practitioners. Participation was strictly voluntary, and reviewers were not compensated for their time. To minimize bias, a blinded evaluation design was employed: reviewers were blinded to the true model identities, with outputs evaluated using fixed, neutral labels (Model A, Model B, Model C) that remained consistent throughout the questionnaire, and the true mapping revealed to the research team only after all scoring was complete; reviewers submitted their scores independently via an anonymous digital questionnaire (Google Forms), so that the authors likewise remained blinded to reviewer identity during data collection and analysis. Responses from MediLITE-QA (merged fp16 checkpoint) and the base Phi-3.5-mini-instruct were generated with HuggingFace Transformers in fp16 on GPU, and o3-mini responses were obtained through the OpenAI API (reasoning effort set to high). All three models received the same question text and were asked for the same output format—a brief explanation followed by a final answer—with minor differences in system-prompt wording; the two Phi-3.5 models used greedy decoding with a limit of 512 new tokens.

To address the limitations of accuracy metrics in capturing hallucination risk, each of the five reviewers independently scored a separate, non-overlapping stratified random sample of 20 questions (proportionally drawn from the 25-item high-risk suite, MedMCQA, and PubMedQA), for a total of 100 distinct question instances evaluated across the panel. For every question in their assigned sample, each reviewer scored all three models (MediLITE-QA, base Phi-3.5-mini-instruct, and o3-mini as a frontier reference) using a four-dimensional rubric: Clinical Correctness (1–5), Harm Severity (0–3), Guideline Adherence (0–2), and Actionability/Triage Quality (0–2). Because each reviewer's 20 items were independently sampled rather than shared across the panel, the unit of paired comparison is the individual (reviewer, question) instance – within each such instance, all three models were rated on the identical clinical scenario by the same reviewer, which is sufficient to support valid paired model-vs-model significance testing across all $n = 100$ instances, though it does not support conventional inter-rater agreement statistics across reviewers (see Section 4.6). Results are reported in Section 3.4.

2.7. Energy, Carbon, and Cold-Start Measurement Protocol

We directly instrumented inference energy for all nine models using CodeCarbon 3.3.0 with root-level Intel RAPL access (`/sys/class/powercap/intel-rapl`, verified before every model) on the dedicated laptop described above, using llama-cpp-python 0.3.35 with the same GGUF Q4_K_M files and inference settings ($n_{\text{gpu_layers}} = 0$, 4 threads) as the latency benchmark in Figure 2. Measurements covered 25 prompts spanning varied clinical task types and lengths (drug interactions, dosing calculations, triage judgments, diagnostic reasoning), with 2 runs per prompt (50 inferences per model and mode), to address prompt sensitivity in inference energy [29,30]. Each model was measured in two

modes. In *fixed length* mode, end-of-sequence and end-of-turn tokens were suppressed so that every run generated exactly 256 tokens, which separates per-token hardware cost from differences in response length. In *natural length* mode, each model stopped on its own. Before measurement, each model completed one warm-up generation, which was discarded, and its idle CPU-package power was then recorded for 30 s with the model loaded (0.8–1.2 W for all models). Consecutive models were separated by a 60 s pause, and the laptop remained on AC power for all measurements. Decoding was greedy, and token counts were taken from llama.cpp's usage report. Each model's chat template was set explicitly (Table 3). An earlier configuration had applied the ChatML template to OpenBioLLM-8B and BioMistral-7B, which caused premature termination and inflated their per-token energy; all reported results use the corrected templates. We report Joules per output token as the primary metric, together with Joules per response, prompt-level bootstrap 95% confidence intervals, and per-prompt paired ratios relative to MediLITE-QA (Section 3.5). The reported Joules are CPU-package RAPL energy (mean power 20.4–23.3 W under load) and exclude DRAM, for which CodeCarbon provides a modelled estimate rather than a measurement, and whole-system (wall-plug) power; explicit reporting of this serving boundary, alongside idle overhead and infrastructure factors, is recommended practice for LLM energy studies [31]. Illustrative gCO_2e figures multiply measured energy by a cited global average grid-intensity factor [32], not the device's actual electricity source, which was not separately metered.

We estimated fine-tuning energy by trapezoidal integration of the 3969 GPU power readings logged in the original training run's Weights & Biases telemetry (`system.gpu.0.powerWatts`; median sampling interval 11.8 min, maximum 2.64 h). The trainer recorded a runtime of 1136.6 h on a single NVIDIA Quadro RTX 4000, and the telemetry spans 1136.1 h of it, consistent with uninterrupted training. This runtime corresponds to approximately 35,000 optimizer steps (three epochs over 373,350 records at an effective batch size of 32), or about two minutes per step for 4-bit QLoRA training with eager attention on this GPU. Power draw was stable across the run (time-weighted mean 122 W, maximum 132 W). Integrating these readings gives 138.75 kWh of GPU energy, or approximately 67 $kgCO_2e$ at the same illustrative global average grid intensity [32]. Intervals longer than 1 h contribute 13.8% of this estimate (assigning them zero energy gives 119.6 kWh), and no formal uncertainty bounds were established. The figure covers the GPU board only (CPU, memory, power-supply losses, and cooling are excluded) and therefore does not represent total training energy.

Separately, because our benchmark host has 6.96 GiB of total RAM, the 7–8B-parameter comparison models' peak memory footprint (Section 3.2) reaches 84–89% of total system memory. To characterize the deployment consequence of this, we separately report each model's *cold-start* TTFT—the very first generation immediately following model load—against its steady-state TTFT (mean of the two subsequent runs), rather than blending both into a single distribution; this comparison, alongside each model's peak memory footprint as a percentage of total device RAM, is presented and interpreted in Section 3.5.

3. Results

3.1. Comparative Analysis and Statistical Significance

Table 4 presents the performance of MediLITE-QA against established baselines. It is important to note that results for proprietary models are sourced from the literature under differing prompting protocols (GPT-4 [33]: 0-shot; Med-PaLM-2 [34]: 5-shot), whereas MediLITE-QA is evaluated under strict zero-shot conditions to reflect the intended deployment setting. Consequently, we classify these comparisons as *indicative* rather than

direct. Table 4 makes this split explicit: MediLITE-QA, base Phi-3.5-mini-instruct, LLaMA 3 8B, OpenBioLLM-8B, BioMistral-7B, and Gemma 2B were evaluated directly (zero-shot, under one identical `lm-evaluation-harness` protocol; Section 2.4), whereas OpenBioLLM-70B, GPT-4, Med-PaLM-2, and Mistral-7B figures are drawn from published literature under different (often 5-shot) prompting protocols and are reported here for indicative context only; OpenBioLLM-70B's 70B-parameter scale in particular makes direct evaluation on our hardware (Section 2) impractical, consistent with its treatment as an indicative literature baseline throughout this paper.

Because MediLITE-QA and its own base model were evaluated under an identical zero-shot protocol with per-item outputs logged by the evaluation harness, we compute a paired significance test for this specific comparison, unlike the literature-sourced baselines: exact McNemar's tests on each benchmark's discordant pairs, Holm-corrected across the nine benchmarks (Table 5). MediLITE-QA scored significantly higher on MedQA (263 items correct only by MediLITE-QA vs. 120 correct only by the base model, out of 1273; $p_{adj} < 10^{-4}$) and MedMCQA (572 vs. 275, out of 4183; $p_{adj} < 10^{-4}$), the two benchmarks with the largest item counts and correspondingly greatest statistical power. The remaining seven benchmarks (Anatomy, Clinical Knowledge, College Biology, College Medicine, Medical Genetics, Professional Medicine, PubMedQA) showed directionally positive or mixed but non-significant differences after correction, consistent with their substantially smaller item counts. We report this as evidence that LoRA fine-tuning produced a statistically detectable accuracy improvement over the base model specifically on the two highest-power benchmarks, rather than claiming an improvement across all nine, which the data do not support at this sample size.

The same per-item protocol was applied to every directly evaluated open model in Table 4, which allows paired comparisons between MediLITE-QA and each of them (all per-benchmark tests are listed in Supplementary Table S4). Averaged over the nine benchmarks (paired item-level bootstrap), MediLITE-QA exceeded its base model by +4.06 pp (95% CI 2.40 to 5.70), did not differ significantly from OpenBioLLM-8B (+0.70 pp, 95% CI −1.29 to 2.65), was modestly higher than LLaMA 3 8B (+2.26 pp, 0.45 to 4.13), and was higher than BioMistral-7B (+16.09 pp, 13.78 to 18.41) and Gemma 2B (+37.69 pp, 35.21 to 40.24). Per-benchmark McNemar tests located MediLITE-QA's advantage over both 8B models on MedMCQA only (OpenBioLLM-8B: +4.71 pp, $p_{adj} < 10^{-4}$; LLaMA 3 8B: +3.99 pp, $p_{adj} < 10^{-4}$), with no significant difference on the remaining eight benchmarks. The average-difference intervals are not adjusted for the five model comparisons; the LLaMA 3 8B interval lies close to zero and should be read as a modest advantage only.

This paired test is not available for the literature-sourced indicative baselines (GPT-4, Med-PaLM-2, Mistral-7B-v0.1), as their reported figures do not include per-item logs; we therefore continue to report those comparisons descriptively, consistent with their treatment as indicative rows in Table 4. Separately, latency results in Figure 2 are descriptive: each model was run three times, the first run is reported separately as the cold-start value (Section 3.5), and steady-state values are the mean of the two subsequent runs; with two warm runs per model we do not report tail percentiles. Formal paired significance testing (Wilcoxon signed-rank and McNemar tests) was also performed for the human-evaluation results in Section 3.4, where per-item outputs from all compared models were directly available.

Table 4. Per-benchmark accuracy (%) for MediLITE-QA against direct (evaluated by the authors, zero-shot) and indicative (literature-reported) baselines. CK = Clinical Knowledge, MG = Medical Genetics, An = Anatomy, PM = Professional Medicine, CB = College Biology, CM = College Medicine, MQA = MedQA (4 options), PQA = PubMedQA, MCQA = MedMCQA. Direct rows were evaluated with `lm-evaluation-harness` using the same 0-shot protocol in fp16 (Section 2.4); the row marked § is the deployed Q4_K_M GGUF file scored on the same items. Averages include item-level bootstrap 95% confidence intervals. Paired tests against MediLITE-QA are given in Section 3.1 and Table 5, and per-benchmark Wilson intervals in Supplementary Table S2. Sources of indicative rows: GPT-4, Nori et al. [33] (0-shot); Med-PaLM-2, Singhal et al. [34] (5-shot; see Section 4.5 on comparability); Mistral-7B-v0.1, Wu et al. [35]; OpenBioLLM-70B, its published model card [36]. Indicative rows were obtained under different protocols and do not support strict ranking against the direct rows. The proposed model, MediLITE-QA, is shown in bold.

Model	Condition	Source	CK	MG	An	PM	CB	CM	MQA	PQA	MCQA	Avg
OpenBioLLM-70B	Indicative	Literature	92.93	93.19	83.90	93.75	93.82	85.74	78.16	78.97	74.01	86.05
Med-PaLM-2	Indicative	Literature (5-shot)	88.30	90.00	77.80	95.20	94.40	80.90	79.70	79.20	71.30	84.08
GPT-4	Indicative	Literature (0-shot)	86.04	91.00	80.00	93.01	95.14	76.88	78.87	75.26	69.52	82.85
LLaMA 3 8B	Direct	0-shot, fp16	72.83	85.00	69.63	70.22	79.17	65.32	59.15	75.00	58.40	70.52 [68.65, 72.37]
OpenBioLLM-8B	Direct	0-shot, fp16	78.11	83.00	68.89	79.04	83.33	68.79	58.13	71.80	57.69	72.09 [70.22, 73.94]
Mistral-7B-v0.1	Indicative	Literature	68.68	71.00	55.56	68.38	68.06	59.54	50.82	75.40	48.20	62.85
Phi-3.5-Mini (base)	Direct	0-shot, fp16	77.36	75.00	65.19	69.85	84.03	66.47	48.55	76.80	55.30	68.73 [66.75, 70.62]
BioMistral-7B	Direct	0-shot, fp16	61.89	67.00	49.63	54.41	61.81	54.91	46.27	71.40	42.96	56.70 [54.55, 58.79]
Gemma 2B	Direct	0-shot, fp16	34.72	27.00	34.81	22.06	36.81	34.68	29.46	66.40	29.88	35.09 [33.05, 37.14]
MediLITE-QA	Direct	0-shot, fp16	**78.87**	**80.00**	**65.19**	**76.84**	**88.19**	**68.21**	**59.78**	**75.60**	**62.40**	**72.79 [70.95, 74.66]**
MediLITE-QA (Q4_K_M) §	Direct	Deployed GGUF, same items	78.87	79.00	66.67	74.26	85.42	72.25	56.40	74.80	59.19	71.87 [70.05, 73.76]

Table 5. Exact McNemar's tests (Holm-corrected across the nine benchmarks) comparing MediLITE-QA against its base model on per-item zero-shot correctness. N is the number of items in the benchmark; MediLITE-only and Base-only are discordant-pair counts (items answered correctly by one model but not the other). * indicates a statistically significant difference after Holm correction ($p_{adj} < 0.05$).

Benchmark	N	MediLITE-Only	Base-Only	p_{adj}	Sig.
Anatomy	135	10	10	1.000	
Clinical Knowledge	265	15	11	1.000	
College Biology	144	9	3	0.876	
College Medicine	173	14	11	1.000	
Medical Genetics	100	10	5	1.000	
MedMCQA	4183	572	275	$<10^{-4}$	*
MedQA (4 options)	1273	263	120	$<10^{-4}$	*
Professional Medicine	272	41	22	0.158	
PubMedQA	500	21	27	1.000	

MediLITE-QA was evaluated on nine medical question answering (QA) benchmarks to assess its performance relative to larger, more resource-intensive language models. With 3.8 billion parameters, MediLITE-QA achieved an average accuracy of 72.79% (95% CI 70.95–74.66) in the fp16 evaluation and 71.87% (70.05–73.76) as the deployed Q4_K_M GGUF file. Among the directly evaluated open models, its average accuracy was not significantly different from OpenBioLLM-8B (72.09%; +0.70 pp, 95% CI −1.29 to 2.65), modestly higher than LLaMA 3 8B (70.52%), and substantially higher than BioMistral-7B (56.70%) and the general-purpose Gemma 2B (35.09%), with the paired tests reported above. Larger models such as GPT-4 (82.85%), Med-PaLM-2 (84.08%), and OpenBioLLM-70B (86.05%) scored higher on average under their own protocols—an expected result given their order-of-magnitude larger parameter counts and, for the two proprietary models, cloud-scale training compute. MediLITE-QA's result therefore shows that domain-specific fine-tuning of a 3.8B model can reach average accuracy not significantly different from that of an 8B domain-specific model on these benchmarks, not that it approaches frontier systems. In terms of efficiency, MediLITE-QA achieved a mean steady-state Time to First Token (TTFT) of 1.34 s and a Time Per Output Token (TPOT) of 0.117 s. Steady-state TTFT was similar across the 7–8B models (1.41–2.06 s), whereas their TPOT was 0.193–0.252 s

(OpenBioLLM-8B: 1.41 s TTFT, 0.206 s TPOT; LLaMA 3 8B: 1.58 s, 0.211 s), and MediLITE-QA's single-request throughput of 8.19 tokens/s was 1.6–2.1× that of the 7–8B models (3.91–5.01 tokens/s; Figure 3). All latency, throughput, memory, and energy measurements used quantized GGUF Q4_K_M models in a CPU-only environment (11th Gen Intel Core i5-1135G7, 6.96 GiB RAM) via llama-cpp-python with GPU acceleration disabled; accuracy was evaluated as described in Section 2.4. On this hardware a typical 150-token answer completes in approximately 19 s (1.34 s + 150 × 0.117 s), which supports interactive response times for non-time-critical information support but not synchronous use during a live consultation.

Table 6 separates the contributions of fine-tuning and quantization (per-benchmark values in Supplementary Tables S2 and S2b). Fine-tuning raised average accuracy by +4.06 pp in fp16 and by +3.62 pp (95% CI 1.97–5.25) between the two Q4_K_M files, so the gain survives quantization. Quantizing MediLITE-QA to Q4_K_M lowered average accuracy by 0.91 pp, a difference whose interval includes zero (−2.04 to +0.22); at the benchmark level, however, the losses on the two largest benchmarks were significant (MedQA −3.38 pp, $p_{adj} = 0.002$; MedMCQA −3.20 pp, $p_{adj} < 10^{-4}$), so quantization is not lossless. Excluding PubMedQA, which supplies 86% of the corpus tokens (Table 2), the fp16 averages are 72.43% (MediLITE-QA) and 67.72% (base): the fine-tuning gain comes from MedQA and MedMCQA, and MediLITE-QA was non-significantly lower than its base model on PubMedQA itself (−1.20 pp).

Table 6. Ablation of fine-tuning and quantization: nine-benchmark average accuracy (%) with item-level bootstrap 95% confidence intervals. Differences are paired item-level bootstrap estimates; the base model's quantization difference was not separately tested.

Model	fp16	Q4_K_M	Quantization Effect
Phi-3.5-mini (base)	68.73 [66.75, 70.62]	68.25 [66.34, 70.15]	−0.48
MediLITE-QA	72.79 [70.95, 74.66]	71.87 [70.05, 73.76]	−0.91 [−2.04, +0.22]
Fine-tuning effect	+4.06 [+2.40, +5.70]	+3.62 [+1.97, +5.25]	

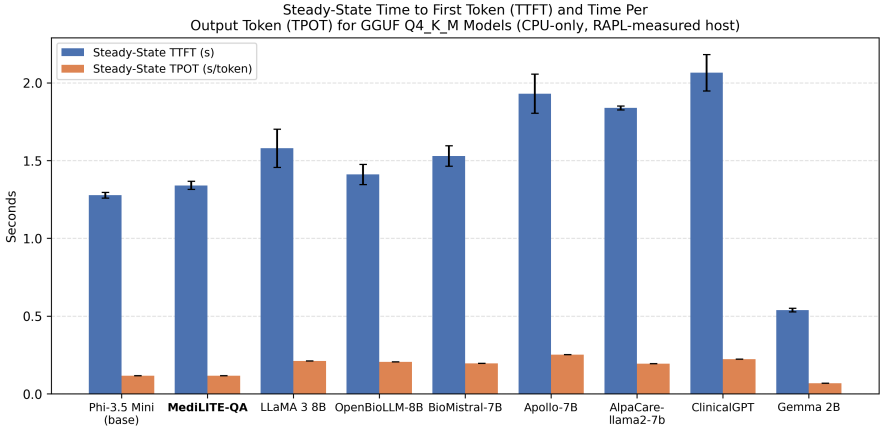


Figure 2. Steady-state Time to First Token (TTFT, s) and Time Per Output Token (TPOT, s/token) across GGUF Q4_K_M quantized models under the CPU-only inference configuration described in Section 2. Values are from the two warm runs following one cold-start run (reported separately in Section 3.5); TTFT error bars show the standard deviation of the two warm runs.

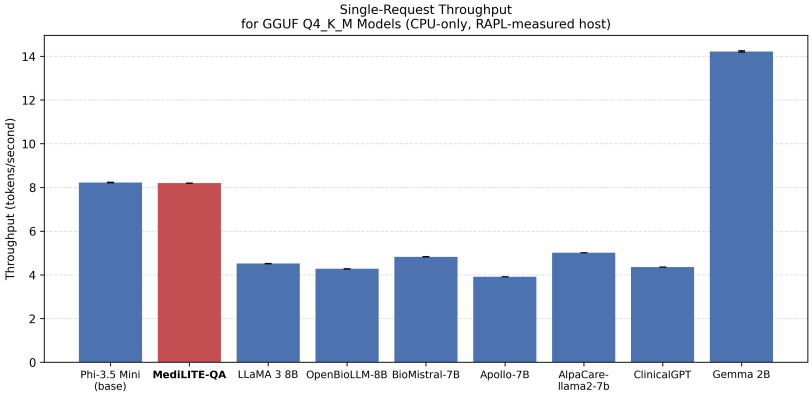


Figure 3. Single-request throughput (tokens/second) across GGUF Q4_K_M quantized models under the same CPU-only inference configuration as Figure 2. Error bars denote standard deviation across runs. AlpaCare-llama2-7b measured 5.01 tokens/second for a single request, slightly below its range of 5.09–5.17 tokens/second over 2–16 sequential requests.

For the full set of GGUF models, Figure 4 reports sustained throughput over 1–16 sequential requests, and Figure 5 plots average accuracy against steady-state Time Per Output Token, which serves as a proxy for relative energy cost per generated token and agrees with the direct energy measurements in Section 3.5. MediLITE-QA stays above 8 tokens/second across 1–16 consecutive requests. Because these requests are issued one after another rather than batched, this result reflects sustained single-stream throughput, not batching efficiency. In the accuracy-latency trade-off, MediLITE-QA occupies the most favorable region observed among the compared models, the lower-right quadrant of Figure 5, combining average accuracy not significantly different from the best directly-evaluated open model (72.79% vs. OpenBioLLM-8B's 72.09%) with a TPOT essentially tied with its own base model at the same Q4_K_M quantization (0.117 s/token for both) and 46–61% of the TPOT of every 7–8B-parameter comparison model. The accuracy gain from LoRA fine-tuning therefore does not come with a latency improvement over the base model, only rough parity with it; MediLITE-QA's advantage over the comparison set is driven by its smaller parameter count and its fine-tuning-derived accuracy gain, not by any latency improvement over its own base model. Figures 6 and 7 show the per-benchmark accuracy for the literature-reported (indicative) baselines and for the directly evaluated models, respectively.

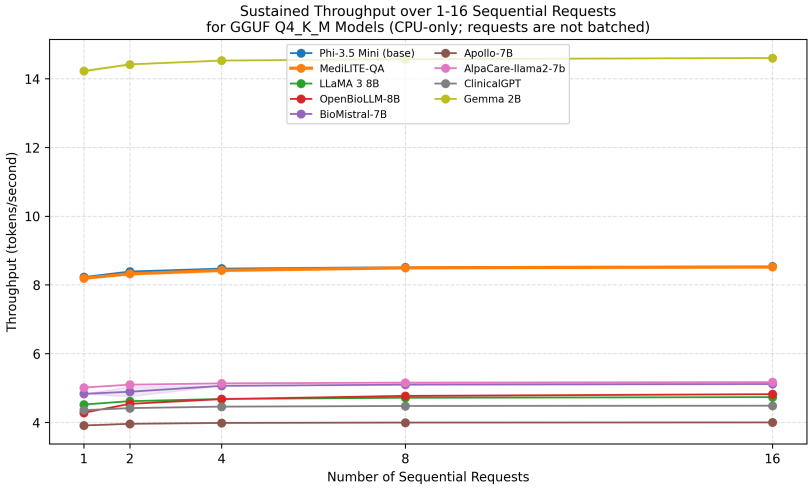


Figure 4. Sustained throughput (tokens/second) over 1–16 consecutive requests for GGUF Q4_K_M quantized models under the same CPU-only inference configuration as Figure 3. Requests are issued sequentially, not batched; shaded bands show the standard deviation across runs.

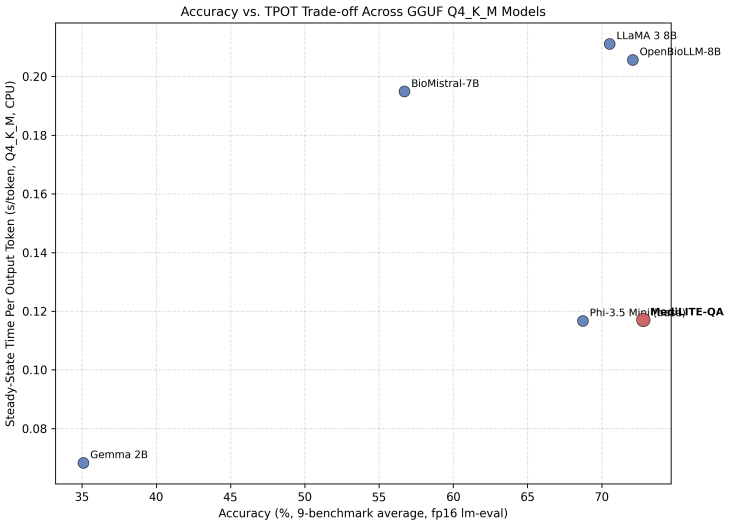


Figure 5. Accuracy vs. steady-state Time Per Output Token (TPOT) trade-off. Accuracy is the nine-benchmark average from the logged fp16 evaluation (Table 4); TPOT is measured on the GGUF Q4_K_M files on CPU. Models without item-level accuracy logs (Apollo-7B, AlpaCare-llama2-7b, ClinicalGPT) are omitted. The lower-right region represents the most favorable combination of high accuracy and low per-token latency.

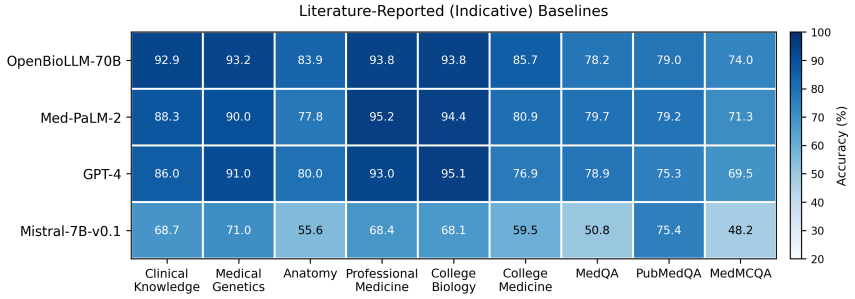


Figure 6. Comparative performance across nine medical benchmarks for literature-reported (indicative) baselines. Color intensity corresponds to accuracy (%), with darker blue indicating a higher score. GPT-4, Med-PaLM-2, and Mistral-7B-v0.1 figures are drawn from published literature under different (often 5-shot) prompting protocols, as detailed in Table 4.

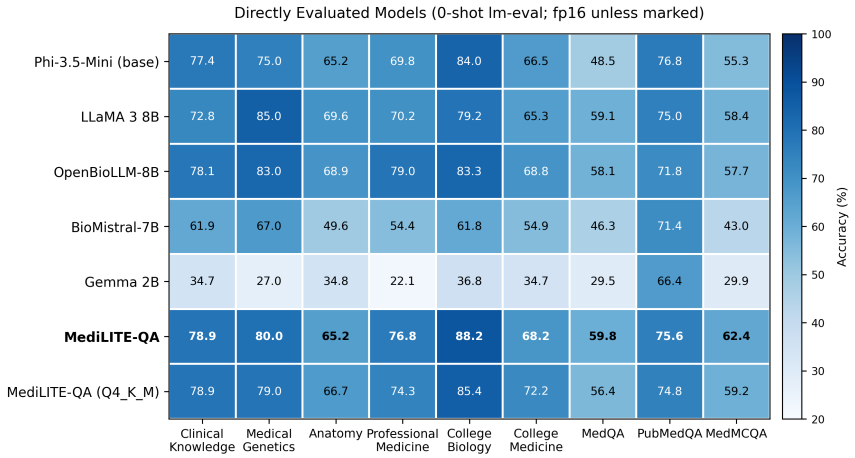


Figure 7. Comparative performance across nine medical benchmarks for models evaluated directly under the same 0-shot `lm-evaluation-harness` protocol in fp16 (MediLITE-QA, base Phi-3.5-mini-instruct, LLaMA 3 8B, OpenBioLLM-8B, BioMistral-7B, and Gemma 2B), plus the deployed MediLITE-QA Q4_K_M file scored on the same items (Section 2.4). Color intensity corresponds to accuracy (%), with darker blue indicating a higher score.

3.2. Memory and Storage Footprint

Table 7 reports the on-disk quantized model size and peak resident memory (RSS) attributable to each model during inference on the dedicated laptop described in Section 2, for the full set of models in Figures 2 and 3. Because this is non-virtualized, dedicated hardware, peak RSS was consistent across repeated runs (standard deviation ≤0.09 GiB per model). We report peak RSS at a 2000-token context window, matching the $N_{ctx} = 2048$ configuration used throughout this paper's inference runs; for Phi-3.5-mini and MediLITE-QA, RSS increased modestly and predictably with context length across the 512–2000-token range tested, whereas the 7–8B models' RSS stayed nearly constant, consistent with being bounded by available physical memory (Section 4.4 discusses the resulting latency implications of this footprint relative to the device's total RAM). OpenBioLLM-8B's Q4_K_M on-disk size (4.58 GB, identical to LLaMA-3-8B's, as both share the same 8B-parameter Llama-3 architecture at the same quantization level) is reported below. All models in this table use Q4_K_M quantization. Smaller on-disk size and lower peak memory both translate directly into lower hardware requirements for deployment—e.g., a device can retain more headroom for other applications, use lower-cost RAM configurations, or, on battery-powered hardware, reduce memory-subsystem power draw—which we revisit as a resource-sustainability indicator in Section 4.4, where we show that for this specific 6.96 GiB device, the 7–8B-parameter models' peak memory footprint reaches 84–89% of total system RAM, with measurable latency consequences.

Table 7. On-disk size and peak resident memory (RSS) at a 2000-token context window during CPU-only GGUF inference, measured on the dedicated laptop described in Section 2. % of Total RAM = peak RSS divided by the device's 6.96 GiB total memory. For the 7–8B models, peak RSS barely changes between 512- and 2000-token contexts (Figure 8), indicating that these values are bounded by available physical memory rather than reflecting each model's unconstrained footprint. RSS is in MiB and total memory in GiB (as reported by the operating system); disk sizes are in GB. The proposed model, MediLITE-QA, is shown in bold.

Model	Disk Size (GB)	Peak RSS @ 2000-Tok ctx (MiB)	% of Total RAM (Peak RSS/6.96 GiB)
Phi-3.5-mini-instruct (base)	2.39	4446	62.4%
MediLITE-QA	2.39	4403	61.8%
LLaMA 3 8B	4.58	6316	88.6%
OpenBioLLM-8B	4.58	6023	84.5%
BioMistral-7B	4.07	6044	84.8%
Apollo-7B	4.96	6280	88.1%
AlpaCare-llama2-7b	3.80	6250	87.7%
ClinicalGPT	4.91	6280	88.1%
Gemma 2B	1.39	2824	39.6%
o3-mini	N/A (cloud-hosted API; no local footprint)		

The GGUF-quantized checkpoint measured above as MediLITE-QA is hosted at `segestic/MediLITE-QA-GGUF` and is provided in four quantization variants (F32, Q8_0, Q5_K_M, Q4_K_M); all results in this paper use the Q4_K_M variant (2.39 GB), matching the quantization level used for every other model in Figures 2 and 3 (accuracy in Table 4 is evaluated in fp16, with the Q4_K_M file reported separately). The corresponding Hugging-Face Transformers-format checkpoint, `segestic/MediliteQA-phi3.5-mini-all-med`, was used for the automated cross-model and human evaluations (Sections 3.3 and 3.4).

3.3. Automated Cross-Model Evaluation

To complement the standard nine-benchmark harness evaluation above and enable direct, paired automated comparison across the three models used in our human evaluation (Section 3.4), we additionally ran MediLITE-QA, base Phi-3.5-mini-instruct, and o3-mini under an identical generation protocol on a separate, stratified 105-question set: a 25-item

custom high-risk clinical-safety suite described in Section 2.5, plus 40 MedMCQA and 40 PubMedQA items. This set is distinct from, and not directly comparable to, the nine-benchmark accuracy figures in Table 4, as it uses a different (smaller, custom) question mix designed to enable controlled three-way comparison rather than to replicate standard benchmark leaderboards.

On the 80 multiple-choice items (MedMCQA + PubMedQA) within this set, exact-match accuracy was 43.8% (35/80) for MediLITE-QA, 40.0% (32/80) for base Phi-3.5-mini, and 58.8% (47/80) for o3-mini. Notably, MediLITE-QA failed to produce a parseable multiple-choice answer in the expected format on 14 of 80 items (17.5%), which were scored as incorrect, whereas the base model and o3-mini produced a parseable answer on all 80 items. Because the base model received the identical prompt, chat template, and decoding settings, this failure is attributable to the fine-tuned model rather than to the evaluation setup: MediLITE-QA was fine-tuned to emit a bare answer letter in a different prompt layout, and the fine-tuning appears to have reduced its adherence to the requested explanation-then-answer format. Among the 66 items it did answer in the expected format, 35 (53.0%) were correct. This loss of instruction following for structured output is a limitation of the present fine-tuning recipe; training on responses in the deployment format is a direct remedy (Section 5).

Across all 105 questions (including open-ended high-risk items), semantic similarity to the reference answers was measured with ROUGE-L and BERTScore (F1, `distilbert-base-uncased`). MediLITE-QA achieved a ROUGE-L of 0.081 and BERTScore of 0.735, compared to 0.069/0.728 for base Phi-3.5-mini and 0.090/0.752 for o3-mini. MediLITE-QA thus scored slightly higher than its untuned base model on both semantic-similarity metrics—small differences that were not tested for significance, in the same direction as the non-significant correctness difference in the human evaluation (Section 3.4)—while trailing the frontier reference model. We note ROUGE-L scores are low in absolute terms across all three models (<0.1), which is expected for open-ended clinical reasoning responses that are semantically appropriate but lexically distinct from the (comparatively terse) reference answers; BERTScore, which captures semantic rather than surface-level overlap, is accordingly the more informative of the two metrics for this task.

3.4. Human Evaluation of Clinical Correctness and Safety

Design and protocol for this evaluation are described in Section 2.6; here we report the resulting scores. Because each reviewer scored a different set of items, inter-rater reliability cannot be estimated, and we therefore treat these results as exploratory (Section 4.6).

Clinical Correctness. Mean correctness scores across the 100 instances were 3.04 (MediLITE-QA), 2.92 (base Phi-3.5-mini), and 3.88 (o3-mini). A Friedman omnibus test confirmed a significant overall difference among the three models ($p < 10^{-4}$). Holm-corrected pairwise Wilcoxon signed-rank tests showed MediLITE-QA scored significantly lower than o3-mini ($p_{adj} = 0.0023$) and the base model also scored significantly lower than o3-mini ($p_{adj} = 5.5 \times 10^{-4}$), while MediLITE-QA vs. its base model was not significant ($p_{adj} = 0.55$). Reviewer-clustered bootstrap 95% confidence intervals for the mean correctness differences are consistent with these tests: MediLITE-QA minus base, +0.120 [−0.120, 0.340]; MediLITE-QA minus o3-mini, −0.840 [−1.080, −0.570] (sensitivity analyses in Section 4.6). With the full 100-instance sample, we therefore find clear statistical evidence that both laptop-deployable models underperform the frontier reference on clinical correctness, but no significant evidence that fine-tuning improved correctness over the untuned base model at this sample size—a finding directly relevant to whether the resource savings characterized in Sections 3.1 and 3.2 currently come at an acceptable accuracy cost (Section 4.1).

Harm Severity and Safety. Using the criterion of any instance receiving a Harm Severity rating ≥ 2, MediLITE-QA was flagged on 7 of 100 instances (7%), the base model on 8 of 100 (8%), and o3-mini on 1 of 100 (1%). Exact McNemar tests (Holm-corrected) showed the base model was flagged significantly more often than o3-mini ($p_{adj} = 0.047$); MediLITE-QA vs. o3-mini approached but did not reach significance ($p_{adj} = 0.063$, 6 discordant instances favoring o3-mini and 0 favoring MediLITE-QA); MediLITE-QA vs. the base model was not significant ($p_{adj} = 1.0$). The base model's difference from o3-mini is therefore borderline significant after correction (8 vs. 1 flagged instances), and MediLITE-QA's is not (7 vs. 1); with flag counts this small, these results do not establish a difference in clinical safety and require validation on a larger sample.

3.5. Energy, Carbon, and Cold-Start Characterization

Table 8 reports CPU-package energy for 25 clinical prompts of varied type and length, with 2 runs per prompt (50 inferences per model and mode), measured as described in Section 2.7. Each model was measured at fixed length, with end-of-sequence tokens suppressed so that every run generated exactly 256 tokens (achieved in 100% of runs for all nine models), and at natural length, where each model stopped on its own; the fixed-length condition separates hardware cost from response length.

Table 8. CPU-package energy (Intel RAPL via CodeCarbon 3.3.0) during CPU-only GGUF Q4_K_M inference over 25 clinical prompts × 2 runs. *Fixed*: exactly 256 generated tokens per run. *Natural*: each model stops on its own. Ratios are computed per prompt relative to MediLITE-QA; bracketed values are prompt-level bootstrap 95% confidence intervals. Power is mean ± SD across runs; idle CPU-package power with the model loaded was 0.8–1.2 W for all models. DRAM energy is excluded because CodeCarbon only models it and does not measure it through RAPL. The proposed model, MediLITE-QA, is shown in bold.

	Fixed Length (256 Tokens)			Natural Length	
Model	**CPU Power (W)**	**J/Token**	**Ratio to MediLITE-QA**	**Mean Tokens**	**J/Response**
Phi-3.5 Mini (base)	23.1 ± 0.8	3.17 [3.15, 3.19]	1.00 [1.00, 1.00]	256	802 [796, 809]
MediLITE-QA	**23.1 ± 0.5**	**3.17 [3.14, 3.19]**	**1.00**	**224**	**713 [630, 783]**
LLaMA 3 8B	23.0 ± 0.7	5.35 [5.33, 5.39]	1.69 [1.68, 1.70]	182	985 [814, 1152]
OpenBioLLM-8B	23.1 ± 0.1	5.33 [5.31, 5.35]	1.68 [1.68, 1.69]	56	343 [279, 428]
BioMistral-7B	23.1 ± 0.2	5.00 [4.98, 5.03]	1.58 [1.58, 1.58]	83	452 [389, 526]
Apollo-7B	23.0 ± 0.2	6.27 [6.24, 6.29]	1.98 [1.97, 1.99]	188	1191 [1004, 1367]
AlpaCare-llama2-7b	22.9 ± 0.3	4.93 [4.91, 4.96]	1.56 [1.55, 1.56]	154	782 [625, 939]
ClinicalGPT	23.3 ± 0.4	5.69 [5.67, 5.72]	1.80 [1.79, 1.80]	140	823 [689, 969]
Gemma 2B	20.4 ± 1.0	1.73 [1.72, 1.73]	0.55 [0.54, 0.55]	256	438 [436, 441]

At matched output length, MediLITE-QA and its base model consume the same energy per token (3.17 J/token; ratio 1.00), as expected from their shared architecture, and every 7–8B-parameter comparison model used 1.56–1.98× more CPU energy per token than either. The two Llama-3-8B models (OpenBioLLM-8B and LLaMA 3 8B) likewise agree within 0.5% (5.33 vs. 5.35 J/token), a consistency check on the measurement. Gemma 2B is the most energy-efficient model per token (0.55× MediLITE-QA) but scores at or near chance on the MMLU subsets (Table 4). Per-response energy additionally depends on response length, which varied four-fold under natural stopping (56–256 tokens): because MediLITE-QA produced longer default responses (mean 224 tokens) than OpenBioLLM-8B (56) or BioMistral-7B (83), its natural-length energy per response (713 J) exceeded theirs (343 and 452 J) despite its lower per-token cost, which makes response-length control a deployment lever in its own right. For an illustrative 256-token response, MediLITE-QA's fixed-length CPU energy corresponds to approximately 0.11 gCO_2e, compared with 0.17–0.21 gCO_2e for the 7–8B comparison models. These values use a global average grid carbon intensity of 480 g CO_2/kWh [32] as an approximation; the actual intensity depends on the electricity

source powering the device, which was not separately metered. They cover the CPU package only and are not a lifecycle comparison with cloud inference. Interpretation of these results in light of MediLITE-QA's accuracy (Section 3.1) is given in Section 4.4.

Table 9 reports each model's cold-start (first generation after load) TTFT against its steady-state TTFT, alongside peak memory footprint as a percentage of the device's total RAM.

Table 9. Cold-start (first-generation-after-load) TTFT versus steady-state (mean of the two subsequent runs) TTFT, alongside peak memory footprint as a percentage of the benchmark device's 6.96 GiB total RAM (from Table 7). Models are ordered by RAM utilization. The proposed model, MediLITE-QA, is shown in bold.

Model	% of Total RAM	Cold-Start TTFT (s)	Steady-State TTFT (s)	Ratio
Gemma 2B	39.6%	0.50	0.54	0.93×
MediLITE-QA	**61.8%**	**1.28**	**1.34**	**0.96**×
Phi-3.5 Mini (base)	62.4%	1.29	1.28	1.01×
OpenBioLLM-8B	84.5%	8.72	1.41	6.18×
BioMistral-7B	84.8%	26.72	1.53	17.48×
AlpaCare-llama2-7b	87.7%	6.64	1.84	3.61×
ClinicalGPT	88.1%	4.76	2.06	2.30×
Apollo-7B	88.1%	16.99	1.93	8.80×
LLaMA 3 8B	88.6%	14.29	1.58	9.05×

All models using 39–62% of total device RAM show a cold-start ratio near 1× (no measurable penalty), whereas all models using 84–89% show ratios of 2.3–17.5×. Within the high-memory group, the penalty does not increase monotonically with memory use (ClinicalGPT: 88.1%, 2.3×; BioMistral-7B: 84.8%, 17.5×), which is consistent with it depending on the operating system's page-cache state at load time. Each cold-start value is a single measurement on one device, so these figures describe this configuration only and do not establish a general memory threshold for CPU-only deployment. Figure 8 plots each model's peak memory footprint against context length, with the device's total RAM shown as a reference line.

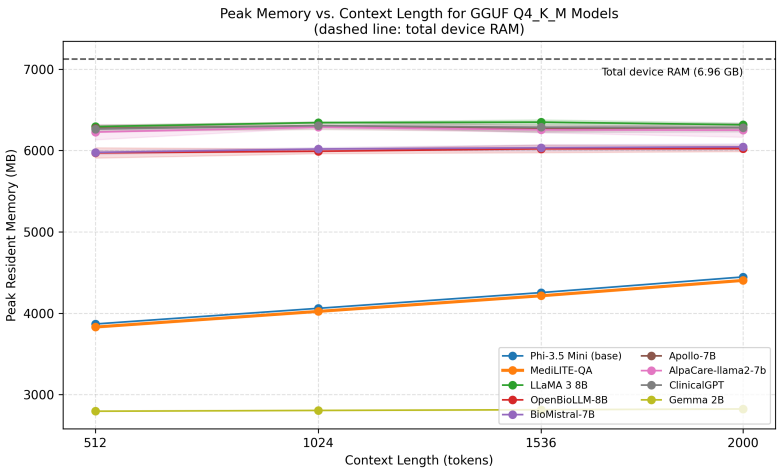


Figure 8. Peak resident memory (RSS) as a function of context length (512–2000 tokens) for GGUF Q4_K_M models on the dedicated laptop described in Section 2, with a dashed reference line marking the device's total 6.96 GiB RAM. Shaded regions denote standard deviation across three runs. This is the visual counterpart to the cold-start finding in Table 9: models whose memory trace sits well clear of the reference line show no cold-start penalty, while models approaching it show a substantial one.

Interpretation of the cold-start pattern and its implications for hardware sizing on this device class is given in Section 4.4.

3.6. Qualitative Error Analysis

To identify recurring failure patterns in MediLITE-QA's responses, we manually reviewed the free-text comments that the five reviewers provided with their scores. The comments fell into three categories.

Safety-critical omission or hallucination on emergency presentations. On several high-risk items, MediLITE-QA correctly identified secondary aspects of a case but missed or contradicted the single most safety-critical action. On a pediatric fever-and-rash presentation, reviewers noted that the model "hallucinated HSP; missed meningococcal septicemia, delaying life-saving IV antibiotics"; on another item, they noted that it "misses neutropenic sepsis emergency," and the model unsafely suggested routine evaluation instead of urgent admission. On a suspected aortic dissection case, one reviewer flagged a "critical safety failure": the model recommended aspirin/heparin, which is appropriate for acute coronary syndrome but contraindicated and potentially fatal in dissection. On a pediatric dosing question, the model was noted to have "hallucinate[d] math and maximum daily limits," risking underdosing. In this pattern, the model recognizes part of a case correctly but gets the critical action wrong or omits it. It was the most clinically consequential failure mode we observed and directly motivates the safety-flag findings in Section 3.4.

Correct final answers with hallucinated mechanistic justifications. In multiple cases, MediLITE-QA chose the correct multiple-choice option but justified it with a fabricated or incorrect causal mechanism. One reviewer noted that the model "selected [the correct option] but justification hallucinates ferroportin for entry," and in another case the model reached the correct answer through reasoning that a reviewer described as "clinically flawed (relied on density, not radiobiology)." This risk differs from that of an outright wrong answer: a clinician who relies on the model's stated rationale, and not only on its final answer, could be misled even when the displayed answer is correct.

Reduced reliability on nuanced yes/no/maybe research-conclusion questions and complex staging/classification items. Reviewers repeatedly flagged incorrect responses on PubMedQA-style questions requiring tolerance of genuine research uncertainty (the "maybe" category), with the model tending to collapse ambiguous findings into an overconfident binary yes/no answer (consistent with the absence of "maybe" labels in its PubMedQA training data; Section 2.1), and on formal clinical staging/classification tasks (e.g., incorrectly assuming a definite T1 tumor stage in a case where staging could not be assessed and the correct classification was Tx). These items require integrating multiple pieces of evidence or explicit handling of stated uncertainty, rather than direct fact retrieval, and were disproportionately represented among incorrect responses across reviewers.

Link to the harm flags. Of the seven MediLITE-QA instances flagged with Harm Severity ≥2 (Section 3.4), five fall in the first category (the pediatric dosing, neutropenic sepsis, aortic dissection, meningococcal septicemia, and refractory asthma items, all from the high-risk suite). The reviewer comments on the remaining two record an incorrect answer without describing its nature, so they could not be assigned to a category. No flagged instance falls in the second category, which by definition involves a correct final answer, and none could be attributed to the third. All seven flagged responses also received the lowest guideline-adherence and actionability scores (0) and a correctness score of 1–2. Four of the seven were also flagged for the base model on the same item, so only three are failures specific to the fine-tuned model, while the base model had four flags that MediLITE-QA did not.

Together, these patterns suggest that MediLITE-QA's current limitations are concentrated less in basic factual recall (where it performs competitively; Section 3.1) and more in high-stakes safety judgment, faithfulness of stated reasoning to the actual answer, and multi-step or uncertainty-aware inference—clarifying the specific respects in which MediLITE-QA

currently requires human-in-the-loop supervision for complex diagnostic tasks, and correspondingly the respects in which resource efficiency alone does not yet make it a sustainable substitute for cloud-hosted clinical decision support in accuracy-critical use.

4. Discussion

4.1. Efficiency, Accuracy, and the Sustainability Trade-Off

The results answer the three research questions in complementary ways. For RQ1, QLoRA fine-tuning produced a significant average accuracy gain over the base model (+4.06 pp in fp16), most of which was retained after Q4_K_M quantization (+3.62 pp), although quantization was not lossless on the two largest benchmarks (Section 3.1). For RQ2, the compact 3.8B model used less CPU energy per token and left substantially more memory headroom than the 7–8B comparison models on the tested laptop, with average benchmark accuracy not significantly different from that of OpenBioLLM-8B (Sections 3.1–3.5). For RQ3, the exploratory clinician evaluation showed no significant correctness gain over the base model and a significant gap to the frontier reference (Section 3.4), so engineering efficiency does not yet translate into clinical usability.

The results show that MediLITE-QA combines domain-specific accuracy with resource efficiency on this hardware. Its average accuracy (72.79% in fp16; 71.87% as the deployed Q4_K_M file) is not significantly different from that of the 8B domain model OpenBioLLM-8B and modestly higher than LLaMA 3 8B, at less than half the parameter count and without the cold-start penalty observed for the 7–8B models, which used 1.56–1.98× more CPU energy per generated token. Its gains over its own base model are concentrated on MedQA and MedMCQA (Section 3.1). A key insight relevant to sustainable AI design is therefore that targeted, parameter-efficient adaptation of a compact base model can reach average benchmark accuracy not significantly different from that of a larger domain model at a lower per-token resource cost, not that it surpasses frontier systems, which remain well ahead. MediLITE-QA's latency and throughput (steady-state TTFT 1.34 s, TPOT 0.117 s/token, 8.19 tokens/s) support interactive, asynchronous use on CPU-only systems—a typical 150-token answer completes in about 19 s—particularly where infrastructure, energy, and privacy are constrained simultaneously. These results characterize one consumer-laptop CPU configuration; future work will measure performance and energy on other device classes, including servers and genuinely mobile/embedded hardware.

We distinguish two claims that the results support to very different degrees. *Engineering deployment feasibility*—that the model runs fully offline on a commodity laptop CPU within its memory budget, at interactive latency and measured energy cost—is directly demonstrated. *Clinical usability*—that its answers are accurate and safe enough to inform care—is not: the exploratory human evaluation (Section 3.4) found MediLITE-QA significantly below a frontier reference on clinical correctness and flagged for potential harm on 7% of instances, and benchmark accuracy of about 72% is not a clinical-grade standard. MediLITE-QA is therefore not a replacement for clinicians or for cloud-hosted clinical decision support; its efficiency gains become a sustainability benefit only once this accuracy and safety gap is closed.

4.2. Privacy and Security Considerations

Local deployment offers a distinct privacy advantage over cloud-centric models by keeping queries on the device, which removes transmission to third-party servers and the associated interception risk. Unlike federated learning, which requires synchronized updates across devices, independent on-device inference is architecturally simpler and keeps data under local control, which also reduces a health system's long-term dependence on any single external cloud provider. Local deployment does, however, shift responsibility

for security to the endpoint. A laptop can be lost, stolen, or compromised by malware; prompts, responses, and cached files stored on it may contain patient information unless disk encryption and no-logging or retention policies are enforced; model files can be tampered with unless their integrity is verified (e.g., by signed checksums); and, without a central service, security patches and model updates depend on local maintenance practices that low-resource settings may struggle to sustain. Model-level threats such as model inversion also remain; future iterations should consider injecting noise during training, as in differentially private training [37], to limit the extraction of sensitive training data from model weights. On-device inference therefore improves privacy only to the extent that the endpoint itself is secured.

4.3. Equity, Accessibility, and Alignment with the Sustainable Development Goals

This work relates to three Sustainable Development Goals. For **SDG 3 (Good Health and Well-Being)**, a model that runs offline on commodity CPU hardware removes the connectivity precondition that excludes many rural and low-infrastructure settings from AI-assisted medical information. For **SDG 9 (Industry, Innovation and Infrastructure)**, it shows that a model with competitive benchmark accuracy can be deployed without accelerator hardware or continuous high-bandwidth connectivity, although its fine-tuning required a GPU (Section 2.7). For **SDG 10 (Reduced Inequalities)**, a CPU-only, offline-capable pipeline lowers (though does not remove) one structural barrier between high- and low-resource settings. These are potential benefits of the deployment model, not measured outcomes of this study, and they are presently offset by the accuracy and safety gap documented in Sections 3.4 and 4.1.

4.4. Energy and Carbon Footprint

Direct energy measurements (Section 2.7, Table 8) quantify the efficiency that latency and memory figures only suggest. At matched output length, MediLITE-QA and its base model use the same CPU energy per token (3.17 J/token), as expected for a shared architecture, and every 7–8B-parameter comparison model uses 1.56–1.98× more. Against OpenBioLLM-8B, whose average accuracy is not significantly different from MediLITE-QA's (Section 3.1), this means MediLITE-QA uses about 40% less CPU energy per generated token (ratio 1.68) with no detectable difference in average benchmark accuracy. A non-significant difference does not by itself establish equivalence, which would require a formal equivalence test. The energy saving comes from the smaller base model: fine-tuning changes accuracy but not inference cost. Gemma 2B has the lowest energy per token (0.55× MediLITE-QA), but its accuracy on the MMLU subsets is near chance.

Response length. Energy per response also depends on response length. With natural stopping, mean response length ranged from 56 tokens (OpenBioLLM-8B) to 256 tokens (Phi-3.5, Gemma 2B). Because MediLITE-QA's default responses were longer (224 tokens), its energy per response (713 J) was higher than that of OpenBioLLM-8B (343 J) and BioMistral-7B (452 J), even though its per-token cost was lower. Per-token energy at fixed length is therefore the appropriate hardware comparison, while per-response energy in deployment additionally depends on response-length policy (e.g., a maximum token budget or a prompt requesting concise answers), which is an energy lever in its own right.

Fine-tuning energy and amortization. The one-time fine-tuning run was estimated to consume approximately 139 kWh of GPU-board energy (Section 2.7). Relative to each 7–8B comparison model, MediLITE-QA saves 452–794 J per 256-token response, so the fine-tuning energy is recovered after approximately 0.63–1.10 million such responses (about 0.90 million relative to OpenBioLLM-8B). Relative to its own base model there is no per-response saving, so the fine-tuning energy is never recovered at inference time; it buys

accuracy, not efficiency. These break-even figures combine a GPU-only training boundary with a CPU-package inference boundary and assume 256-token responses, so they are illustrative orders of magnitude rather than precise accounting.

Scope of the sustainability claim. All energy figures here are measured local CPU inference on one device. We did not measure cloud inference energy, and a lifecycle comparison with cloud deployment would additionally require data-center overheads, network transfer, and embodied hardware emissions on both sides; our results therefore support a comparison among models on the same local hardware, not a claim that local inference is lower-carbon than cloud inference in general.

Cold-Start Latency and Memory-Constrained Deployment

The cold-start pattern reported in Section 3.5 (Table 9) reflects the RAM requirements of CPU-only deployment at this model scale on this device: every model using 39–62% of total device RAM shows a cold-start ratio near 1×, while every model using 84–89% shows a single-run ratio of 2.3–17.5×. Within the high-memory group, the penalty does not increase monotonically with memory use, which is consistent with swap and page-cache pressure that varies with the operating system's cache state when a multi-gigabyte checkpoint is loaded into a system with little free memory. In a separate diagnostic run on this device, a representative 7B model used 3.7–4.1 GB of swap, and this usage persisted during inference, not only at load time. Its steady-state TPOT (Figure 2) nonetheless stayed within the range expected for its parameter count, so the main measurable effect is on the first query. This matters for the low-infrastructure deployments considered in this paper: a quantized model file that fits on a device's storage does not show that the device has enough RAM headroom to run it without a first-query latency penalty, so hardware sizing guidance for CPU-only deployment of 7B-parameter-class models should state RAM headroom explicitly, not only storage footprint. At 3.8B parameters, MediLITE-QA stays well below this memory range on this device. This is a practical, secondary benefit of fine-tuning a smaller base model for the domain, separate from its accuracy and per-token energy advantages. Because these observations come from a single device and one cold-start measurement per model, they should not be generalized to other RAM sizes or CPU architectures without further measurement. Figure 8 (Section 3.5) shows the same relationship: across the 512–2000-token range tested, the smaller models (Phi-3.5-Mini, MediLITE-QA, Gemma 2B) stay well below the device's RAM ceiling, whereas the 7–8B-parameter models come close enough to it that longer contexts would leave little headroom.

4.5. Clinical Readiness and Limitations

MediLITE-QA is currently a research prototype, not a certified medical device. Before any clinical use, it would require external validation on local patient populations, regulatory approval (e.g., under the FDA Software as a Medical Device guidelines), and a medico-legal liability assessment. Our results should be read as a proof of concept for the engineering feasibility of resource-light, on-device medical AI, not as a recommendation for clinical deployment. Additionally, MediLITE-QA is currently trained and evaluated exclusively on English-language medical text; its applicability to non-English clinical settings—including the multilingual, often lower-resource settings most relevant to the equity motivation in Section 4.3—is untested, and multilingual extension (Section 5) would require both additional training data and dedicated evaluation rather than a direct transfer of the present results. Fairness across populations and regions also could not be assessed: the evaluation datasets do not provide consistent individual-level demographic or geographic annotations, so subgroup performance cannot be validly estimated, and we do not infer fairness across groups from aggregate benchmark accuracy. The source datasets

draw largely on Indian medical entrance examinations (MedMCQA), US licensing examinations (MedQA), and English-language biomedical abstracts (PubMedQA), which may under-represent the clinical presentations, disease burden, and practice guidelines of other regions; population- and region-stratified evaluation is required before any claim of equitable performance.

A further limitation concerns the comparability of the accuracy figures in Table 4. While MediLITE-QA and the directly-evaluated open-source baselines were assessed using a strict 0-shot pipeline, the comparative scores reported for proprietary and literature-sourced models (GPT-4 [33], Med-PaLM-2 [34], and the Mistral-7B and OpenBioLLM-70B figures drawn from external benchmark reporting) reflect different, often more favorable, prompting configurations—the Med-PaLM-2 figures, for instance, use 5-shot prompting [34], and few-shot prompting of this kind can meaningfully improve reported scores relative to a 0-shot baseline. Because prompting protocol, not only model capability, differs across these rows, direct head-to-head ranking between our 0-shot results and these literature baselines should be read as contextual rather than strictly controlled; MediLITE-QA's accuracy relative to the directly-evaluated open-source models (Table 4, all under the same 0-shot protocol) is the more methodologically comparable result. The independent 13-gram contamination screen (Section 2.4) found verbatim overlap for 0.57% of evaluated items and left all conclusions unchanged when those items were excluded; it cannot, however, detect paraphrased overlap, which an embedding-based screen would be needed to rule out.

4.6. Human Evaluation Limitations

The human evaluation design has one central limitation that should be considered when interpreting the results in Section 3.4. Because each of the five reviewers scored an independently drawn, non-overlapping 20-question sample rather than a shared fixed panel of items, conventional inter-rater reliability statistics (e.g., Fleiss' Kappa, intraclass correlation) cannot be computed: these metrics require multiple reviewers to rate the identical item, which did not occur here. We are therefore unable to report how consistently different reviewers would have scored the same clinical scenario, and cannot rule out that rubric interpretation varied across the panel. The model-vs-model comparisons in Section 3.4 remain statistically valid because, within each reviewer's own set of items, all three models were rated on the identical question by the same reviewer, supporting paired significance testing across the full $n = 100$ reviewer-question instances; what is not supported is any claim about agreement *between* reviewers. In addition, the human and automated cross-model evaluations used the fp16 merged checkpoint (Sections 2.6 and 3.2), not the deployed Q4_K_M file, whose benchmark accuracy is 0.91 pp lower on average (Table 6). Because the model labels (A, B, C) were fixed across items, reviewers could in principle have learned to associate a label with a model's characteristic style, which limits the strength of the blinding. Future iterations of this evaluation should use a shared, fixed item panel across all reviewers specifically to enable inter-rater reliability estimation alongside the model comparison, and should include a calibration round in which reviewers jointly score a small pilot set before the blinded evaluation begins to reduce rubric interpretation variability. For this reason, and because of the small sample, the human-evaluation results in Section 3.4 should be treated as exploratory evidence of relative model performance, not as a validated safety or quality assessment; a follow-up study with a shared reviewer panel is included in future work.

The 100 correctness observations are also clustered within only five reviewers (20 instances per reviewer), so we ran two reviewer-level sensitivity analyses to check whether the MediLITE-QA comparisons in Section 3.4 depend on any single reviewer. First, in a leave-one-reviewer-out analysis, we removed each reviewer's 20 instances in turn and recomputed

the MediLITE-QA-vs-base mean correctness difference. The difference stayed positive in all five cases (range: +0.025 to +0.213), so no single reviewer determines the direction of this comparison. Second, a reviewer-clustered bootstrap (10,000 resamples of the five reviewers, each resampled with its full 20-item block to preserve within-reviewer correlation) gave a point estimate of +0.120 for MediLITE-QA vs. base (95% CI: $[-0.120, 0.340]$), consistent with the non-significant Wilcoxon result in Section 3.4, and −0.840 for MediLITE-QA vs. o3-mini (95% CI: $[-1.080, -0.570]$), consistent with the significant Wilcoxon result. The clustered confidence intervals are wider than a naive instance-level interval would be, as expected given the five-reviewer clustering, but do not change either comparison's qualitative conclusion. With only five clusters, however, bootstrap intervals are themselves imprecise and tend to be too narrow, so they should be read as approximate.

4.7. Adaptation Strategies Beyond LoRA

MediLITE-QA adapts its base model with LoRA, which requires a labelled domain corpus and a one-time training run (estimated at approximately 139 kWh of GPU-board energy in this study; Section 2.7). Where labelled medical data are scarce, or where the target setting—a clinical specialty or a language – differs from the fine-tuning corpus, prompt-based parameter-efficient methods offer a lighter-weight alternative: prompt tuning [23] and prefix tuning [24] keep the base weights frozen and learn only a small set of continuous prompt vectors. Cai et al. [25] show how far this idea can be taken in a specialised medical task. Their MedLP-HAFB-CLIP model adapts a pretrained vision–language model (CLIP) to the identification of second-trimester fetal ultrasound standard planes through a medical prompt learner, whose learnable context vectors generate prompts matched to anatomical categories, enriched with fine-grained anatomical descriptions and an auxiliary loss for planes that sonographers find easy to confuse; a hierarchical adaptive feature block extracts and adaptively fuses multi-scale image features.

Although its modality and task differ from text-based question answering, MedLP-HAFB-CLIP offers three methodological analogies for the present pipeline. First, learnable prompts that encode clinical specialty or language context could adapt a single LoRA-tuned model to new subdomains, or to the multilingual settings identified in Section 4.5, without updating or merging the base-model weights. Second, its injection of structured domain knowledge into the prompts suggests a route to the staging and classification failures observed in Section 3.6. Third, its hierarchical design suggests layering adaptation—a general medical LoRA adapter combined with lightweight specialty-specific prompts—rather than training a separate adapter for each subdomain. These strategies carry a deployment cost specific to CPU-only inference, however: a LoRA adapter can be merged into the base weights and quantized into a single GGUF file with no inference overhead, whereas learned soft prompts are prepended to every query, lengthening the input the CPU must process before the first token and so raising TTFT and per-query energy, and they require embedding-level input that standard GGUF chat interfaces do not expose. Whether prompt-based or hierarchical adaptation complements the present pipeline is therefore an empirical question about this accuracy–efficiency trade-off, which the measurement protocol used here is designed to quantify (Section 5).

5. Conclusions

This study presents MediLITE-QA, a lightweight, domain-specific model fine-tuned for medical question answering and deployed CPU-only on a consumer laptop. Built on Phi-3.5-mini-instruct with LoRA (QLoRA) fine-tuning and 4-bit GGUF quantization, the deployed model reaches 71.87% average accuracy across nine medical QA benchmarks (72.79% before quantization, not significantly different from the 8B domain model OpenBioLLM-8B),

with a peak resident memory of 61.8% of total RAM and no cold-start penalty, while the 7–8B comparison models used 1.56–1.98× more CPU energy per generated token. These findings establish engineering feasibility on this consumer-laptop CPU tier; they do not establish clinical usability. The exploratory blinded clinical evaluation found no significant correctness gain over the base model and a significant gap to a frontier cloud reference, and the fine-tuning stage itself was estimated to consume approximately 139 kWh of GPU-board energy, recovered only after roughly 0.6–1.1 million responses relative to larger models. MediLITE-QA should accordingly be read as evidence that compact, domain-specific medical LLMs are a worthwhile direction for lower-infrastructure healthcare AI relevant to SDG 3, 9, and 10, and as a measured baseline for future iterations, rather than as a ready-to-deploy clinical or environmental sustainability solution.

Future Work

To bridge the gap between this research prototype and both clinical translation and a fully substantiated sustainability case, future work will focus on six concrete directions:

- **Shared-Panel Clinical Evaluation:** a follow-up human evaluation in which all reviewers score a common item set, enabling inter-rater reliability estimation, with a larger sample and a calibration round.
- **Format-Preserving Fine-Tuning:** training on responses in the intended deployment format (explanation followed by a final answer) rather than bare answer letters, to retain the base model's structured-output instruction following (Section 3.3).
- **Prompt-Based and Hierarchical Adaptation Beyond LoRA:** following Section 4.7, future work will compare prompt tuning [23], prefix tuning [24], and hierarchical prompt-based adaptation inspired by MedLP-HAFB-CLIP [25] against the present LoRA pipeline for specialty- and language-specific adaptation. The comparison will measure not only accuracy but also the additional TTFT and per-query energy introduced by soft prompts on the same CPU hardware.
- **Retrieval-Augmented Generation (RAG):** integrating a local vector database of current clinical guidelines (e.g., CDC, WHO) to reduce hallucinations and allow knowledge updates without full model retraining, itself a lower-energy alternative to periodic re-fine-tuning.
- **Multilingual Adaptation:** extending the instruction-tuning dataset using multilingual corpora (e.g., MMedBench [38]) to support non-English-speaking clinics in the Global South, directly extending the equity aims of Section 4.3.
- **Genuinely Mobile/Embedded Deployment and RAM-Constrained Validation:** extending the present consumer-laptop-tier characterization to smartphones, single-board computers, and other genuinely low-power hardware, with matched energy instrumentation, and specifically testing whether the cold-start/RAM-headroom relationship identified in Section 4.4 (Table 9) generalizes to devices with even less available memory than the 6.96 GiB laptop used here.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/su182111245/s1, Table S1: The 25-item custom high-risk clinical-safety evaluation suite (Section 2.5), with the reference answer for each question; Table S2: Per-benchmark accuracy (%) with Wilson 95% confidence intervals for all directly evaluated models (fp16); Table S2b: Per-benchmark accuracy (%) with Wilson 95% confidence intervals for the deployed Q4_K_M GGUF files; Table S3: Contamination screen: word 13-gram overlap between each evaluation set and the complete fine-tuning corpus; Table S3b: Fine-tuning effect after excluding flagged items; Table S3c: The 25 most frequent 13-grams shared between evaluation items and the fine-tuning corpus (Section 2.4); Table S4: Per-benchmark paired comparisons on identical items (Section 3.1).

Author Contributions: Conceptualization, O.O.O. and S.B.; methodology, O.O.O.; software, O.O.O.; validation, O.O.O. and S.B.; formal analysis, O.O.O.; investigation, O.O.O.; resources, O.O.O.; data curation, O.O.O.; writing—original draft preparation, O.O.O.; writing—review and editing, S.B.; visualization, O.O.O.; supervision, S.B.; project administration, S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study, including the blinded five-reviewer clinical evaluation described in Section 2.6, was approved by the Near East University Scientific Research Ethics Committee (Bilimsel Araştırmalar Etik Kurulu) under application number NEU/AS/2026/302, approved 25 March 2026.

Informed Consent Statement: Informed consent was obtained from all five reviewers involved in the human evaluation study, using the consent-form language presented at the start of the reviewer questionnaire (Section 2.6).

Data Availability Statement: The LoRA adapter is available at https://huggingface.co/segestic/phi3.5-mini-4k-qlora-medical-seg-vall_med (accessed on 25 October 2026) (used for the fp16 accuracy evaluation); the fine-tuned model with this adapter merged into the base weights is available at https://huggingface.co/segestic/MediliteQA-phi3.5-mini-all-med (accessed on 25 October 2026) (HuggingFace Transformers format; used for the automated cross-model and human evaluations) and https://huggingface.co/segestic/MediLITE-QA-GGUF (accessed on 25 October 2026) (GGUF format, available in F32, Q8_0, Q5_K_M, and Q4_K_M quantization variants; all CPU inference results in this paper use the Q4_K_M variant). The training corpus is compiled from the publicly available MedMCQA (`openlifescienceai/medmcqa`), MedQA-USMLE-4-options (`GBaker/MedQA-USMLE-4-options`), and PubMedQA (`bigbio/pubmed_qa`) corpora; the 25-item custom high-risk evaluation suite (Section 2.5) is provided in the Supplementary Materials. Training, evaluation, benchmarking, and energy-measurement code with their configurations, together with the per-item evaluation logs and per-run energy measurements underlying all reported results, are available at https://github.com/segestic/Medilite-QA (accessed on 25 October 2026). Public datasets are referenced rather than redistributed.

Acknowledgments: Computational resources used for model fine-tuning were provided informally by the authors' home department at Near East University.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Hyperparameter and Training Configuration Details

Table A1 consolidates the fine-tuning and hardware configuration parameters referenced throughout Section 2.

Table A1. Fine-tuning and inference configuration.

Parameter	Value
Base model	`microsoft/Phi-3.5-mini-instruct` (3.8B parameters)
Fine-tuning method	QLoRA (LoRA adapters on a 4-bit NF4-quantized base, double quantization)
LoRA target modules	q, k, v, o, gate, up, and down projections
LoRA rank (r)	32
LoRA scaling (α)	256
LoRA dropout	0.10
Optimizer	Paged 8-bit AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Weight decay	0.1

Table A1. *Cont.*

Parameter	Value
Peak learning rate	2×10^{-4}
LR scheduler	Linear decay
LR warmup	100 steps
Batch size	4 per device × 8 gradient-accumulation steps (effective 32)
Maximum sequence length	512 tokens
Training epochs	3
Precision	fp16 mixed precision (the training GPU does not support bf16)
Training hardware	1× NVIDIA Quadro RTX 4000
Training duration and energy	1136.6 h; ≈139 kWh GPU energy (trapezoidal integration of sampled W&B power telemetry); ≈67 $kgCO_2e$ at 480 g/kWh
Validation	5% of the training corpus held out; evaluated every 100 steps; seed 1234
Training corpus size	104.5M tokens, 393,000 records (99.2M used for gradient updates, 5.2M held out for validation; Table 2)
Post-training quantization	GGUF, Q4_K_M (4-bit)
Inference engine	llama-cpp-python (llama.cpp)
Inference hardware	11th Gen Intel Core i5-1135G7 (Tiger Lake), 6.96 GiB RAM (dedicated non-virtualized laptop, 4 physical cores/8 threads via 2-way hyperthreading)
Inference threads	4
GPU layers	0 (CPU-only)

References

1. Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; Huang, K. Mobile Edge Intelligence for Large Language Models: A Contemporary Survey. *IEEE Commun. Surv. Tutor.* **2025**, *27*, 3820–3860. [CrossRef]
2. Zheng, Y.; Chen, Y.; Qian, B.; Shi, X.; Shu, Y.; Chen, J. A Review on Edge Large Language Models: Design, Execution, and Applications. *ACM Comput. Surv.* **2025**, *57*, 1–35. [CrossRef]
3. Friha, O.; Ferrag, M.A.; Kantarci, B.; Cakmak, B.; Ozgun, A.; Ghoualmi-Zine, N. LLM-Based Edge Intelligence: A Comprehensive Survey on Architectures, Applications, Security and Trustworthiness. *IEEE Open J. Commun. Soc.* **2024**, *5*, 5799–5856. [CrossRef]
4. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25–29 April 2022.
5. Sparrenberg, L.; Deußer, T.; Berger, A.; Sifa, R. Small and Fast LLMs on Commodity Hardware: Post-Training Quantization in llama.cpp. In *Proceedings of the 2025 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*; IEEE: New York, NY, USA, 2025. [CrossRef]
6. Wu, C.; Lin, W.; Zhang, X.; Zhang, Y.; Xie, W.; Wang, Y. PMC-LLaMA: Toward building open-source language models for medicine. *J. Am. Med. Inform. Assoc.* **2024**, *31*, 1833–1843. [CrossRef]
7. Busch, F.; Hoffmann, L.; Rueger, C.; Van Dijk, E.H.C.; Kader, R.; Ortiz-Prado, E.; Makowski, M.R.; Saba, L.; Hadamitzky, M.; Kather, J.N.; et al. Current applications and challenges in large language models for patient care: A systematic review. *Commun. Med.* **2025**, *5*, 26. [CrossRef]
8. Kim, H.; Hwang, H.; Lee, J.; Park, S.; Kim, D.; Lee, T.; Yoon, C.; Sohn, J.; Park, J.; Reykhart, O.; et al. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digit. Med.* **2025**, *8*, 240. [CrossRef]
9. Zhang, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Chen, G.; Li, J.; Wu, X.; Zhang, Z.; Xiao, Q.; et al. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates Inc.: Red Hook, NY, USA, 2017.

11. Strubell, E.; Ganesh, A.; McCallum, A. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3645–3650.
12. Rohanian, O.; Nouriborji, M.; Kouchaki, S.; Clifton, D.A. On the effectiveness of compact biomedical transformers. *Bioinformatics* **2023**, *39*, btad103. [CrossRef]
13. Huang, K.; Singh, A.; Chen, S.; Moseley, E.; Deng, C.Y.; George, N.; Lindvall, C. Clinical XLNet: Modeling Sequential Clinical Notes and Predicting Prolonged Mechanical Ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020. [CrossRef]
14. Naseem, U.; Khushi, M.; Reddy, V.; Rajendran, S.; Razzak, I.; Kim, J. BioALBERT: A Simple and Effective Pre-trained Language Model for Biomedical Named Entity Recognition. In *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*; IEEE: New York, NY, USA, 2021; pp. 1–7. [CrossRef]
15. Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief. Bioinform.* **2022**, *23*, bbac409. [CrossRef]
16. Kanakarajan, K.R.; Kundumani, B.; Sankarasubbu, M. BioELECTRA: Pre-trained Biomedical text Encoder using Discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021. [CrossRef]
17. Lu, Q.; Dou, D.; Nguyen, T. ClinicalT5: A Generative Language Model for Clinical Text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022. [CrossRef]
18. 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*, e40895. [CrossRef]
19. Yang, S.; Zhao, H.; Zhu, S.; Zhou, G.; Xu, H.; Jia, Y.; Zan, H. Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback and Real-World Multi-Turn Dialogue. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 19368–19376. [CrossRef]
20. Peng, C.; Yang, X.; Chen, A.; Smith, K.E.; PourNejatian, N.; Costa, A.B.; Martin, C.; Flores, M.G.; Zhang, Y.; Magoc, T.; et al. A Study of Generative Large Language Model for Medical Research and Healthcare. *npj Digit. Med.* **2023**, *6*, 210. [CrossRef]
21. Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; Dufour, R. BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 5848–5864. [CrossRef]
22. Pal, A.; Sankarasubbu, M. OpenBioLLM-8B: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. Hugging Face Repository. 2024. Available online: https://huggingface.co/aaditya/Llama3-OpenBioLLM-8B (accessed on 11 September 2026).
23. Lester, B.; Al-Rfou, R.; Constant, N. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 3045–3059.
24. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 4582–4597.
25. Cai, J.; Dai, C.; Xiong, R.; Wang, B.; Luo, Y.; Zheng, L.; Ma, D. MedLP-HAFB-CLIP: Hierarchical Adaptive Large Model With Learnable Medical Prompts for Level II Ultrasound Standard Plane Identification. *Ultrasound Med. Biol.* **2026**, *52*, 2064–2079. [CrossRef]
26. Rajkomar, A.; Hardt, M.; Howell, M.D.; Corrado, G.; Chin, M.H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [CrossRef]
27. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*; Curran Associates: Red Hook, NY, USA, 2020; pp. 1877–1901.
28. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
29. Poddar, S.; Koley, P.; Misra, J.; Ganguly, N.; Ghosh, S. Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 12688–12704. [CrossRef]
30. Fernandez, J.; Na, C.; Tiwari, V.; Bisk, Y.; Luccioni, S.; Strubell, E. Energy Considerations of Large Language Model Inference and Efficiency Optimizations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 32556–32569. [CrossRef]

31. Hoxha, J.; Thanasi-Boçe, M.; Khalifa, T. A Deployment-Aware Framework for Carbon- and Water-Efficient LLM Serving. *Sustainability* **2025**, *17*, 10473. [CrossRef]
32. International Energy Agency. *Emissions Factors 2023*; IEA: Paris, France, 2023. Available online: https://www.iea.org/data-and-statistics/data-product/emissions-factors-2023 (accessed on 9 September 2026).
33. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. Microsoft Research. 2023. Available online: https://www.microsoft.com/en-us/research/uploads/prod/2023/03/GPT-4_medical_benchmarks.pdf (accessed on 9 September 2026).
34. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [CrossRef]
35. Wu, C.; Qiu, P.; Liu, J.; Gu, H.; Li, N.; Zhang, Y.; Wang, Y.; Xie, W. Towards evaluating and building versatile large language models for medicine. *npj Digit. Med.* **2025**, *8*, 58. [CrossRef]
36. Pal, A.; Sankarasubbu, M. OpenBioLLM-70B: Advancing Open-Source Large Language Models for Healthcare and Life Sciences. Hugging Face Repository. 2024. Available online: https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B (accessed on 11 September 2026).
37. Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H.B.; Mironov, I.; Talwar, K.; Zhang, L. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 308–318. [CrossRef]
38. Qiu, P.; Wu, C.; Zhang, X.; Lin, W.; Wang, H.; Zhang, Y.; Wang, Y.; Xie, W. Towards Building Multilingual Language Model for Medicine. *Nat. Commun.* **2024**, *15*, 8384. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.