African Journal of Education Science and Technology (AJEST) Vol. 8 No. 4 (May, 2026)
University of Eldoret, Kenya, Mount Kenya University, Kenya, Chukwuemeka Odumegwu Ojukwu University, Nigeria, Kyambogo University, Uganda and University of Makeni, Sierra Leone

Deploying Medical Large Language Models on Consumer-Grade Hardware: A Scoping Review of Quantization, Parameter-Efficient Fine-Tuning, and Inference Latency

Kevin Kipmutai[1], Dancun Nyale[1] & Stanley Kiplagat Rotich[2]
[1]Co-operative University of Kenya, Department of Computer Science and Information Technology
[2]Machakos University, Department of Mathematics & Statistics

Corresponding Email: kevinkipmutai@gmail.com

Abstract

There is a challenge of limited access to high-performance computing infrastructure, unreliable electricity, and intermittent internet connectivity in deployment of large language models (LLM) in clinical settings in low- and middle-income countries (LMICs). This scoping review aims to compile evidence regarding the use of medical LLMs on consumer-grade devices using quantization and parameter-efficient fine-tuning (PEFT). Based on the PRISMA extension for scoping reviews (PRISMA-ScR), we conducted searches across PubMed Central, medRxiv, arXiv, and Google Scholar for studies published between 2024 and 2026 on quantized or PEFT-adapted LLMs with reported hardware, memory, latency or accuracy values for medical tasks. Eight studies were found that fulfilled the inclusion criteria, covering clinical data extraction, biomedical question answering, radiology report generation, diagnosis in intensive care, medical text simplification, and diagnostic support in East Africa. Deployments on consumer GPUs like the 3-billion-parameter RTX 3060 (12 GB) and even an 8 GB laptop delivered response times below 4 seconds, while one offline diagnostic system on 3-billion-parameters achieved 100% top-3 diagnostic accuracy on a set of curated clinical cases from East Africa. To test for clinical safety, a benchmark of 26 quantized models used for intensive care showed that most of the models failed a simple patient-safety test, although they responded quickly and with competence. Computational feasibility does not measure clinical safety. Evidence from Africa and other LMIC contexts is still limited, with just one of eight studies having taken place in a sub-Saharan African context, and no studies reporting a completed prospective clinical deployment. The results indicate that quantizing and PEFT enable medical LLMs to be practically used in resource-limited settings, yet clinical validation, safety assessment, and evidence specific to LMICs are still significant challenges.

Keywords: *Quantization, QLoRA, consumer GPU, medical LLM, low-resource setting*

Introduction

The advent of AI into healthcare has grown rapidly, and large language models (LLMs) have shown promising potential in clinical decision making, literature synthesis and patient communication (Bommasani et al., 2021; Moor et al., 2023). These tools become part of the daily clinical routine, and increasing numbers of clinicians require practical guidance on evaluating AI-based studies and understanding the fundamentals of AI (Meskó & Görög, 2020). Some models like GPT-4, Med-PaLM 2, and LLama family models have been performing at levels similar to human doctors in standardized medical exams, thus leading to high hopes for clinical use (Singhal et al., 2023; Tu et al., 2024). The computational demands of these models, however, create significant challenges to their adoption in resource-poor environments, especially in low- and middle-income countries (LMICs) where power supplies are inconsistent, internet bandwidth is limited and high-performance computing equipment is scarce (Wahl et al., 2018; Al-Ganad et al., 2026).

Sub-Saharan Africa has a critical shortage of healthcare workers with less than one physician for every 1,000 people in most countries, and AI driven clinical tools may help fill a gap in diagnostic capacity and standardize clinical judgment (Wahl et al., 2018). However, the industry standard of LLM deployment is based on cloud-based inference on data center GPUs with 40 to 80 GB of VRAM, which is cost prohibitive for most healthcare centers in LMICs (Al-Ganad et al., 2026). The price of a single NVIDIA A100 GPU is higher than the annual budget of many district hospitals in Africa and Asia, and cloud API subscriptions are far too costly to be used as a clinical tool over the long term (Ugarte-Gil et al., 2020; Valli et al., 2026).

This advancement in LLM capability has been largely enabled by the scaling of model size and training compute, which has led to increasingly large models that require increasingly more compute resources for training, fine tuning, and serving

(Kaplan et al., 2020). There are some recent advances in model compression and efficient adaptation that have variable appeal. Quantization techniques quantize model weights from 32-bit floating point to 8-bit integers or 4-bit representations, significantly lowering memory usage and allowing for inference on CPUs and GPUs from typical consumer devices (Frantar et al., 2023; Lin et al., 2024). Parameter-efficient fine-tuning (PEFT) techniques, such as Low-Rank Adaptation (LoRA) and its quantized version QLoRA, enable domain-specific adaptation by modifying only a small subset of parameters, which decreases the memory usage for training by up to 90% from full fine-tuning (Hu et al., 2022; Dettmers et al., 2023). These techniques extend on previous memory-efficient distributed training strategies, including ZeRO that split the optimizer states, gradients, and parameters between devices to make training of very large models feasible without quantization (Rajbhandari et al., 2020). By pairing 4-bit quantization with QLoRA, it has been found that 65 billion parameter models can be fine-tuned on a single 48 GB GPU, implying that 8 billion parameter models like Llama 3 might be fine-tuned on GPUs with less than 16 GB VRAM (Dettmers et al., 2023).

Even with this progress, the pros and cons of computational performance and clinical performance have been insufficiently studied for deployment in resource-limited settings. In medical applications where accuracy is crucial, approximation errors due to quantization could accumulate. The hallucination rate due to aggressive quantization has not been well understood, yet it poses serious challenges in clinical applications (Lee et al., 2023). Additionally, the latency of inferences directly affects clinical usability, with a latency of less than 30 tokens/second considered slow by end-users, but benchmarks of quantized medical LLMs on consumer GPUs are still fragmented (Kwon et al., 2023).

To fill these gaps, this scoping review aims to map the recent peer-reviewed and preprint literature on three interrelated themes: (1) quantization methods for reducing the memory footprint of medical LLMs, particularly focusing on 4-bit and 8-bit precision; (2) parameter-efficient fine-tuning through QLoRA for domain adaptation on consumer-grade GPUs; and (3) the trade-offs between inference latency, clinical accuracy, and safety of medical LLMs in resource-constrained settings. As per the objectives of a scoping review, the purpose is to map the scope, nature and gaps in this rapidly growing evidence base, not a quantitative pooling analysis. Reporting is guided by the PRISMA extension for scoping reviews (PRISMA-ScR; Tricco et al., 2018), and the review is limited to very small models that could be practically implemented on a consumer grade hardware that is widely available in LMICs.

Methodology

This study was performed as a scoping review, using Arksey and O'Malley's methodological framework for scoping reviews and reporting in line with the PRISMA extension for reporting of scoping reviews (PRISMA-ScR; Tricco et al., 2018). The decision to use a scoping review was intentional: the purpose is not to answer a specific effectiveness question by means of pooled statistical synthesis, but to map the size, scope, and characteristics of a very new and growing evidence base, and to define gaps. As per Scoping-Review methodology, formal risk-of-bias scoring and meta-analysis were not performed. The review was not registered prospectively and the search was carried out by one reviewer. These limitations are clearly noted here and are repeated in the discussion.

Search Strategy and Data Sources

Databases for this review were selected based on their coverage of medical literature and preprints relevant to the rapid field of LLM deployment. PubMed Central (PMC) and Google Scholar provide comprehensive coverage of peer-reviewed medical research and allied disciplines. medRxiv and arXiv were included because the field of quantized medical LLMs is rapidly evolving with significant research appearing first as preprints, and inclusion of preprints is explicitly supported by PRISMA-ScR guidelines for emerging topics (Tricco et al., 2018).

Table 1. Search Strategy and Data Sources

Database	Complete Search String/Strategy with Boolean Operators	Records Retrieved
PubMed Central	(("quantization" OR "quantized" OR "4-bit" OR "8-bit" OR "NF4" OR "GGUF" OR "GPTQ" OR "AWQ") OR ("LoRA" OR "QLoRA" OR "PEFT" OR	127

	"parameter-efficient fine-tuning")) AND ("medical" OR "clinical" OR "biomedical") AND ("large language model" OR "LLM") Date range: 2024/01/01 to 2026/07/31 Search date: 2026-08-15	
medRxiv	("quantization" OR "quantized" OR "4-bit" OR "8-bit") AND ("LLM" OR "large language model" OR "NLP") Date range: Posted 2024-01-01 to 2026-07-31 Search date: 2026-08-14	43
arXiv	Submitted Date: [202401010000 TO 202607312359] AND (all:"quantization" OR all:"QLoRA" OR all:"LoRA") AND (all:"medical" OR all:"clinical" OR all:"healthcare") Search date: 2026-08-15	89
Google Scholar	Iterative searches with boolean operators: • ("quantized" OR "quantization") AND ("medical LLM" OR "clinical LLM") • ("QLoRA" OR "LoRA") AND ("medical" OR "clinical") • ("consumer GPU" OR "edge deployment") AND ("medical" OR "healthcare") Date range: 2024-2026; First 100 results per query Search date: 2026-08-15	156

Database Exclusions

There are major reasons Scopus and Web of Science were not included in this review: (1) These are broad, multidisciplinary databases that would increase the search volume without proportionally improving recall for the specific technical-medical intersection of quantized medical LLMs,(2) Most literature focused on quantized LLMs can be found in arXiv (technical preprints) and medRxiv (medical literature) before being indexed in broader databases, making arXiv and medRxiv more efficient primary sources for this new field,(3) PubMed is specifically tailored towards the indexing of literature in the medical domain and it is more comprehensive with respect to peer-reviewed field - specific medical literature compared to Scopus, (4) For single-reviewer scoping review of a rapidly evolving field, content-richness of the search source is balanced by the feasibility of the search, such as PubMed, medRxiv, or arXiv for the technical-medical field intersection of quantized medical LLMs, or Google Scholar for general purpose literature with a lean domain-specific search. This is in line with guidance provided by PRISMA-ScR that states pragmatic database selection is acceptable for scoping reviews. As a caveat, studies that are not listed in this review could be retrieved by researchers who have access to subscription databases, such as Scopus, Web of Science, Embase or IEEE Xplore.

Full searches were undertaken according to the search strategies of Table 1 over four sources from 1 January 2024 to 31 July 2026. The search dates have been chosen to include the most recent publications about quantized medical LLMs, which is a very recent research area (the majority of publications in this area are within these two years). Searches were conducted on August 14-15, 2026 and were all completed.

Search Results Summary

Total records that were identified across all databases: 415. Number of duplicates that were removed: 103. Post deduplication records: 312. Records screened (title/abstract): 312. Records that were excluded after screening: 265. Full-text articles that were assessed for eligibility: 47. Articles excluded after full-text review was done: 39. Included studies in the final review: 8 (5 peer-reviewed publications + 3 preprints).

Inclusion and Exclusion Criteria

The studies included were those that reported original empirical evaluation of a medical or clinical task using either a quantized (8-bit, 4-bit, or GGUF-format) large language model or a PEFT-adapted (LoRA/QLoRA) large language model, and they reported at least one of the following: GPU memory or VRAM usage, inference latency, or an accuracy or quality metric against a full-precision or non-quantized baseline. Studies that focused on general-domain (non-medical) natural language processing (NLP) tasks, the perplexity evaluation of quantization without a clinical outcome, and review papers and position papers without original data were excluded.

Study Selection and Data Extraction

Titles and abstracts returned by each search were sifted by one reviewer for relevance, and candidates were then checked against the full text (or their detailed abstract and available preprint version if full text was behind a paywall) to ensure they were relevant and to extract the relevant data (model, quantization method, hardware, memory, latency, accuracy and task). The charting was descriptive, as is common in a scoping review and did not rely on a formal instrument to appraise the quality of the articles.

Single-Reviewer Screening

One reviewer undertook the data extraction and data screening in this scoping review. With the use of single-reviewer screening, there is a greater risk for selection bias than with dual-independent screening. Single-reviewer screening may be acceptable in exploratory or scoping reviews. However, where the literature may be a rapidly changing body of work. In the current review several reasons justified this: (1) This is a very new field, as evidenced by the very new medical LLM quantization literature, with the majority of studies from 2024-2026. (2) Inclusion criteria are quite technical and would need someone knowledgeable in both the medical field and machine learning to be able to evaluate them, which would likely be pretty time consuming for the dual reviewers. However, these decisions do have limitations that may cause selection bias. Thus, the eight studies in the review can only be seen as a snapshot of, or representative example of, an evolving body of evidence and cannot be assumed to be an exhaustive list of studies that met the inclusion criteria. In a partial mitigation effort, studies were iteratively searched in an attempt to discover potentially missed studies by citation chaining. Future systematic or update reviews on this topic are highly recommended to use dual independent screening to achieve more robust evidence synthesis.

Database Coverage and Study Count

The number of studies included is relatively small (n=8), which is perhaps an indicator of the infancy of the field and inclusion criteria. The field of quantized medical LLMs is only 2-3 years old and there are limited rigorous published evaluation with necessary metrics (hardware configuration, memory, latency, and accuracy outcomes specific to medical fields). Some databases were not included, namely Embase, IEEE Xplore, Scopus and Web of Science, to examine potential search gaps.

Embase was not included due to a large overlap with PubMed in medical literature and its little extra value in the field of pharmaceutical research, which is not directly related to LLM use. The works in IEEE Xplore were excluded as hardware optimisation, quantization, and inference techniques are well described in the technical preprint literature (arXiv), which was selected as a main source for this type of research. The choice of using PubMed, medRxiv, arXiv, and Google Scholar is simply a matter of choosing the most comprehensive set that can be searched given the limitations of a single reviewer. It means however, that reports of eight studies shown here in these four sources represent only those studies that met criteria and did not represent a complete search of all possible technical and biomedical databases. The researchers who can access the subscription databases (Scopus, Web of Science and Embase, IEEE Xplore) may discover other studies that are not captured by this review. It should be noted that most of the most active researchers in quantized LLMs publish their works on arXiv first, and the researchers of medical LLMs publish their works on medrxiv, so the main studies on this fast-changing area are likely captured.

Data Charting and Synthesis

Due to the small number and heterogeneity of the included studies (differing model families, tasks, datasets and hardware), a formal quality-appraisal instrument and meta-analysis were not applied. Instead, each study's design (peer-reviewed publication versus preprint, single-center versus multi-model benchmark, sample size) is reported alongside its results in Table 2 so that readers can weigh the evidence accordingly. The findings are combined and presented in a narrative fashion, categorised into quantization and memory footprint, parameter-efficient fine-tuning, inference latency on consumer hardware, and safety or reliability consideration.

Evidence Quality and Study Heterogeneity

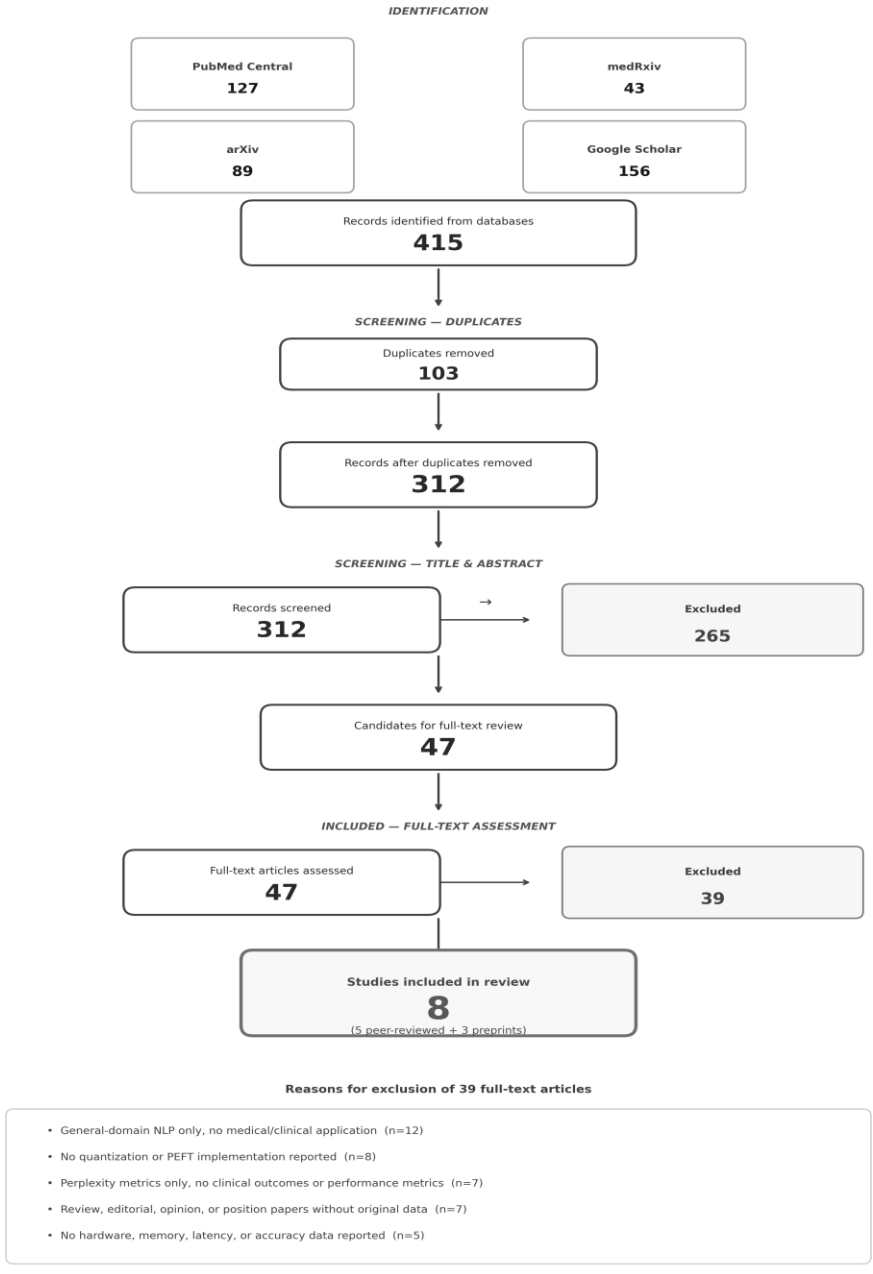
While not formal risk of bias analysis is performed in scoping reviews, it is important to note that studies included offer evidence of varying quality and strength. This is not a fault; it is rather a natural attribute of a newly developing discipline. The eight studies included range from large-scale benchmarks (e.g., Shlyakhta's evaluation of 26 quantized models and their safety profiles) to single-institution pilots (e.g., Walusimbi et al.'s proof-of-concept diagnostic system in Uganda). Five studies were peer-reviewed and published in peer-reviewed journals or conference proceedings, and three preprints (D'addario 2025, Shlyakhta 2026, and Walusimbi et al. 2026) have not yet been formally peer-reviewed. The designs of the studies differ widely ranging from studies with controlled comparisons of quantization methods, on the same models and hardware (Zhan et al., Shlyakhta), to single deployment studies in specific clinical contexts (Walusimbi et al.), and finally to studies benchmarking multiple models against a reference (Kim et al.). These results have been presented in subsequent sections according to the study design and review status, so the reader may consider evidence strength in evaluating the results obtained for each study. Each study has been identified as to whether or not it was peer-reviewed, as stated in Table 2.

Preprint Inclusion and Interpretation

Three of the eight studies surveyed (D'addario 2025, Shlyakhta 2026, Walusimbi et al. 2026) are not yet peer-reviewed and have been posted on preprint servers medRxiv and arXiv. Preprints are the front lines of research in this fast-moving field and are explicitly encouraged for emerging fields and topics by PRISMA-ScR guidelines, but readers should keep in mind that they have not been peer reviewed and could undergo revision or retraction. Preprints are included because: 1) The field is rapidly expanding, and many of the most important, and new, findings should be included in any evidence synthesis as soon as they emerge, which in some cases could be delayed by 6-12 months in the peer-review process,2) The field is recognised as emerging or nascent, and PRISMA-ScR explicitly allows evidence to be included by preprint so as to reflect the new and fast evolving nature of the field, (3) evidence synthesis would be significantly restricted by excluding preprints, and important contributions would be left out of the synthesis (e.g., the comprehensive safety benchmarking by Shlyakhta and Walusimbi et al.'s first-of-a-kind clinical deployment in Uganda), both preprints at the time of writing. The findings from preprint studies should be interpreted with the understanding that they are preliminary. Specifically, the safety benchmarking

data by Shlyakhta (e.g., 78% of models failed on this simple allergy-recall test) are important early safety data which need replication and peer review before informing policy intervention; Walusimbi et al.'s Aletheia system are an important proof of concept but were not extensively performed by clinicians (2 clinicians, uncontrolled pilot) and should not be seen as a validation of a clinical system. The peer-review status of these preprints should be indicated in future updates of this review.

Figure 1: PRISMA-ScR Study Selection Flow Diagram



Results and Discussion

Study Selection and Characteristics

Eight studies fulfilled the inclusion criteria (Table 2). All were published or posted from 2024 to 2026, aligning with the fact that quantized medical LLMs deployment is a very recent research field. This review is based on five peer-reviewed journal publications and three preprints (medRxiv or arXiv) not yet peer-reviewed, as indicated throughout this review. The studies were from the United States, Romania, Germany, Slovakia, Brazil, Uganda and a collaboration between Saudi Arabia and India. Included tasks included clinical data extraction, general biomedical NLP benchmarking, radiology report generation, radiology diagnostic reasoning, intensive care decision support and safety, medical text simplification and offline differential diagnosis support. The sizes of models varied from 3 to 72 billion parameters, with most studies concentrating on the smallest size (8 billion or less) to make the model accessible on consumer hardware.

Table 2. Characteristics of Included Studies

Study	Year	Model(s)	Quantization / PEFT	Hardware	Clinical Task
Shakya et al.	2025*	Llama-3.1-8B-Instruct	LoRA vs. QLoRA (8-bit, 4-bit)	Multi-GPU (count varied)	Clinical data extraction
Zhan et al.	2025*	12 LLMs incl. ClinicalCamel-70B, Meditron-70B, Qwen2.5, Phi-4	8-bit / 4-bit (W8A16, W4A16)	4× A100 40GB	NER, relation extraction, classification, QA
Voinea et al.	2024	Llama 3-8B	4-bit + LoRA (rank 16)	Consumer-grade GPU	Radiology report generation
Kim et al.	2025	15 open-source LLMs (best: Llama-3-70B) + GPT-4o	Local quantized (GGUF, Q4/Q5_K_M)	Local/consumer inference	Radiology differential diagnosis
Shlyakhta	2026*	26 LLMs (best: Granite 3.1/3.2 8B)	Q4_K_M (4-bit GGUF)	RTX 3060, 12GB	ICU decision support & safety
D'addario	2025*	Gemma 3 (1B–27B)	4-bit (Q4_K_M) + QLoRA	Consumer-class VRAM budgets	Medical exam QA (HealthQA-BR)
Walusimbi et al.	2026	Qwen2.5-3B-Instruct	QLoRA + GGUF quantization	8GB laptop, <4GB RAM	Offline differential diagnosis (Uganda)
Srinivasu et al.	2026	Llama-3.1-8B-Instruct	QLoRA (rank 4)	Single consumer GPU	Clinical text simplification

* Preprint at time of writing; not yet certified by peer review.

Quantization Techniques and Memory Footprint

Various quantization techniques were employed in included studies, including 4-bit NormalFloat (NF4) or a variant of it (Q4_K_M) in GGUF format, 8-bit weight quantization (W8A16), post-training quantization (e.g., GPTQ, AWQ), and SmoothQuant (Xiao et al., 2023). The quantization scheme of NF4 has been found by Dettmers et al. (2023) to be information-theoretically suited for weights of normally-distributed neural networks, and is used in the majority of the medical applications discussed here, such as the fine-tuning performed by Shakya et al. (2025), Walusimbi et al. (2026) and Srinivasu et al. (2026), as well as the local inference by Shlyakhta (2026) and Kim et al. (2025) using the GGUF Q4_K_M quantization. Voinea et al. (2024) used radiology report generation as a representative task for the medical domain, where their base model is called Llama 3-8B, which is one of the most widely used families of models in the medical field in this literature.

Table 3 presents the results on memory and accuracy from four of the included studies reporting on direct comparisons of full precision versus quantized comparisons. D'addario (2025) found that the model Gemma 3 12B was quantized into 4-bit and inference VRAM reduced from 98.4 GB to 30.6 GB (a 69% reduction), while fine-tuning VRAM was reduced from 214.3 GB to 43.6 GB (an 80% reduction), with accuracy dropping by only 1.3 percentage points on a Brazilian medical licensing exam and was essentially unchanged (99.7% retained) on a second exam. In the same way, Zhan et al. (2025) reported a similar drop in memory from approximately 28.7 GB to 10.9 GB when evaluating the 14-billion-parameter Phi-4 model on MedQA, with 2.0 percentage points loss in accuracy. At the lower end, Walusimbi et al. (2026) successfully compress a 3-billion-parameter model into less than 2 GB of storage and 4 GB of RAM on a relatively standard 8 GB laptop, demonstrating the capacity to shift medical LLM inference from GPU to CPU-based, laptop-powered workstations.

Table 3. Selected Memory and Accuracy Outcomes from Included Studies

Study	Model	Memory Change (4-bit vs. full precision)	Reported Accuracy/Quality Change
D'addario (2025)	Gemma 3 12B	Inference: 98.4→30.6 GB (−69%); fine-tuning: 214.3→43.6 GB (−80%)	Revalida exam 66.3%→65.0% (−1.3 pp); ENARE exam 99.7% of original accuracy retained
Zhan et al. (2025)	Phi-4 (14B)	MedQA task: 28.67→10.94 GB	MedQA accuracy 64.8%→62.8% (−2.0 pp)
Shakya et al. (2025)	Llama-3.1-8B	QLoRA (4-bit) used about two-thirds of LoRA's peak GPU RAM	Fine-tuning gain: +10–20 pts (LoRA) vs. +8–14 pts (QLoRA) over base model
Walusimbi et al. (2026)	Qwen2.5-3B	Deployed model <2GB; <4GB RAM on an 8GB laptop	Top-1 diagnostic accuracy 80.0%; Top-3 accuracy 100.0%

Parameter-Efficient Fine-Tuning with QLoRA

The most prevalent PEFT method in the studies included in this review was QLoRA (Dettmers et al., 2023) which involved 4-bit quantization of the frozen base model and trainable low-rank adapters (Hu et al., 2022). Although the reported adapter ranks varied from task to task, Voinea et al. (2024) were able to report rank 16 for the task of generating radiology reports, while Srinivasu et al. (2026) were able to report a very low rank of 4 for clinical text simplification using Llama-3.1-8B-Instruct and reported only marginal additional gains when increasing the rank to 8 or 16, which they attributed to the increased amount of parameters. Srinivasu et al. (2026) also analysed adapter placements (a lightweight model that only includes query/value projections, a standard model that includes a key projection as well, and a larger "TinyAttention" model including output projections in every attention layer), and concluded that this placement matters more than the number of parameters in this generation task.

On clinical data extraction, Shakya et al. (2025) observed that QLoRA fine-tuning of Llama-3.1-8B-Instruct enhanced their extraction metrics by 8 to 14 points over the base model, whereas full-precision LoRA raised the scores by 10–20 points, with only 2 to 4 points difference between the two models despite significantly reduced compute requirements. In the context

of medical text simplification, Srinivasu et al. (2026) achieved a Flesch-Kincaid grade level of 7.6 on the Med-EASi dataset, which falls within the range of 6-8 for patient-oriented texts, and a BERTScore of 0.845, showing that there was significant preservation of meaning along with increased readability. Walusimbi et al. (2026) implemented QLoRA fine-tuning on a 3-billion-parameter model with 27,000 curated clinical reasoning examples to yield a top-1 diagnostic accuracy of 80.0% and a top-3 accuracy of 100.0% across 10 different clinical case categories, with the majority representing diseases common in East Africa.

Another practical benefit observed in multiple studies is that the size of the QLoRA adapter weights is small compared to the size of the full model checkpoints, which, in principle, makes it easier to distribute domain-specific medical adapters across limited bandwidth connections.However, none of the studies included measured the logistics of transferring or updating the adapters in terms of bandwidth or memory usage in a real deployment with low bandwidth. So, this is an observation rather than a direct finding from the studies presented here (Hu et al., 2022).

Inference Latency and Real-Time Clinical Usability

Inference latency is a primary determinant of clinical usability for interactive medical AI systems. Only one included study, directly compared the response latency of multiple quantized models on a single fixed consumer device, which is Shlyakhta (2026) as reported in Table 4, which were tested on an NVIDIA RTX 3060 (12 GB VRAM, around $400 at 2026 prices) using 4-bit (Q4_K_M) quantized weights through llama.cpp.

Table 4. Consumer-Hardware Inference Latency (Shlyakhta, 2026)

Model (4-bit, Q4_K_M)	Hardware	Median Latency	Source
Granite 3.2 8B	RTX 3060, 12GB	0.58 s	Shlyakhta (2026)
Llama 3.2 3B Instruct	RTX 3060, 12GB	0.75 s	Shlyakhta (2026)
Phi 3.5 Mini Instruct	RTX 3060, 12GB	0.89 s	Shlyakhta (2026)
Granite 3.1 8B	RTX 3060, 12GB	3.84 s	Shlyakhta (2026)

The median response time in that benchmark varied from 0.58 seconds for the fastest models to several seconds in the slower models, and 12 of 23 (52.2%) of the models tested were measured as fast or moderate (less than 5 seconds). Notably, the two models that were best performing on a clinical safety test in the same study, Granite 3.2 8B and Granite 3.1 8B, had median latencies of 0.58 and 3.84 seconds respectively, and the study found no significant correlation between response speed and safety performance (Pearson r = 0.12), meaning that fast inference on consumer hardware does not come at the cost of safety at least for the models tested.

Other included studies reported latency only indirectly. Zhan et al. (2025) reported inference latency for four models (DeepSeek-LLM-65B, Llama-3.3-70B, Phi-4, and Qwen2.5-72B) run on enterprise-class A100 GPUs rather than consumer hardware, so their absolute latency figures are not directly comparable to Table 4.However, they consistently observed that quantization increased latency relative to full precision, a result they attribute to performing computation in full precision internally to preserve accuracy, which adds a precision-conversion overhead (Zhan et al., 2025). The trade-off, between higher loading and lower memory, with a sacrifice to per-query speed, is in line with other works in the quantization literature (Kwon et al., 2023). Walusimbi et al. (2026) noted that their quantized model fit within the memory budget of an ordinary laptop but failed to give a specific response time, making it difficult to compare the latencies across the set of included studies.

Clinical Accuracy and Safety Considerations

While the preservation of clinical accuracy when quantizing is the most important factor for medical deployment, only a subset of the included studies compared the full precision and quantized model on a common task, the rest compared quantized models to each other or to a human or a proprietary-model baseline.
Kim et al. (2025) compared 15 open-source LLMs, most of which are served locally using quantized (GGUF Q4/Q5_K_M) weights, to the proprietary GPT-4o on 1,933 radiology case reports from the Eurorad library, with correctness defined as the

true diagnosis appearing in the model's top three suggestions. GPT-4o was the best overall model (79.6%), followed closely by the quantized and locally deployable model Llama-3-70B (73.2%), with both models performing close to two experienced radiologists on a smaller local validation set of 60 MRI scans of the human brain. It implies that if careful attention is paid to model selection, open, quantizable models can match the accuracy of proprietary cloud-models in radiology differential diagnosis, but it is not a complete match.

Zhan et al. (2025) is the only included study that directly quantifies the impact of quantization on domain-specialized medical models. They compared the full-precision, 8-bit, and 4-bit versions of ClinicalCamel-70B, HuatuoGPT-o1-70B, Llama3-Med42-70B, and Meditron-70B, with minimal impact reported on quantization on tasks like drug-drug interaction classification and cancer-hallmark labelling, particularly on 4-bit quantization, but with more significant degradation reported at 8-bit precision for HuatuoGPT-o1-70B and Llama3-Med42-70B, pointing to non-uniform effects of quantization across models and bit-widths. In a separate radiology text-generation task, Voinea et al. (2024) reported that their 4-bit, LoRA-adapted Llama 3-8B model produced conclusions preferred over a human radiologist's in roughly 22% of paired comparisons, with independent radiologist reviewers rating the AI output as clinically relevant, though the study did not include a full-precision comparison arm to isolate the effect of quantization specifically.

In the case of Shlyakhta (2026), which tested models for its dual-safety test with 26 quantized models to support ICU decisions on documented drug allergies, the vast majority (18 of 23 models tested, 78.3%) failed to recall a drug allergy and refuse an order for an unsafe drug, even while most models seemed to be fluent and medically knowledgeable. Notably, eight models that successfully withstood a distinct battery of negative hypothetical instructions (an adapted version of the Milgram-obedience test) did not meet the allergy-recall test; and there was only a weak, slightly negative correlation between the two safety measures ($r = -0.39$), implying that abstract ethical refusal and concrete clinical memory are unrelated and can be separated. Due to the variability of the model architecture and training rather than the quantization level, the conclusions of this study do not pertain to quantization-induced hallucination specifically, but rather to the general safety of consumer-hardware-deployable LLMs. No study in this analysis separated out the hallucination rate by bit-width alone.

Implications for Low- and Middle-Income Countries

The evidence collected here is directly applicable, albeit in a limited way, to the deployment of medical AI in LMICs. The best available proof of concept to date is the Aletheia system developed by Walusimbi et al. (2026), which is entirely offline and fine-tuned to disease patterns that are more prevalent in East Africa, and achieved 100% top-3 diagnostic accuracy on curated categories of disease cases. On a broader scale, Shlyakhta (2026) showed that 26 quantized models could be deployed on a single consumer GPU (RTX 3060, costing approximately $400 in 2026), which is an affordable option for university groups and mid-size hospitals in many LMICs (Al-Ganad et al., 2026; Wahl et al., 2018). But even if technically feasible, it's not necessarily clinically ready. On the very same page, even the Aletheia team is clear that their first clinical evaluation, carried out with two clinicians they co-authored with in Uganda, was small-scale, and hasn't been undertaken in the context of a formal study protocol – the next envisioned phase is not finished, but rather is planned for in the future: with several hundred patient cases, across multiple sites, with a formal study protocol (Walusimbi et al., 2026. More generally, Shlyakhta (2026) revealed that 91.3% of the models they tested were insecure for the basic safety of the patient, which is yet another reminder that fluent models, which are run locally, quickly and easily by consumers, are not necessarily safe models. Waller et al. (2024) and the WHO guidelines on AI in health highlight the requirements for medical AI systems to demonstrate higher standards of evidence, such as prospective validation and ongoing monitoring (World Health Organization, 2021). None of the eight studies in this study reported a completed clinical prospective validation and only one study, Walusimbi et al. (2026), reported some real clinician interaction with the system.

Specific evidence in the context of LMIC and Africa is limited. Of the eight studies that were included, only the one by Walusimbi et al. (2026) was specifically in a sub-Saharan African setting; the other six were conducted in the United States, Europe, or other multinational research collaborations, with no particular LMIC focus, and D'addario (2025) addressed a large middle-income health system (the SUS in Brazil) using a Portuguese-language benchmark. This aligns with previous findings that institutions of LMICs and in particular Africa are significantly under-represented in healthcare LLM research compared to disease burden and population (Al-Ganad et al., 2026).

Policy-wise, these studies provide grounds for cautious optimism that local AI infrastructure is more promising than relying solely on cloud APIs, but do not yet provide grounds for specific procurement or regulatory standards, as none of the studies

included model drift, long-term reliability, or outcomes beyond initial benchmarks or small-scale pilots. Regionally validating locally representative clinical data, as Walusimbi et al. (2026) started in Uganda, seems to be the most promising near-term step, along with the continued safety benchmarking exemplified by Shlyakhta (2026) on ordinary consumer hardware.

Conclusion and Recommendations

This scoping review identifies an emerging body of evidence and concludes that it is technically possible to deploy LLMs to consumer-level hardware.: For the eight studies mapped, 4-bit quantization–across all models–reduced GPU memory usage by ~69-80% for inference and fine-tuning, while the accuracy costs seen on medical benchmarks were typically 1-2% for larger models and 2-4% for QLoRA versus full-precision fine-tuning for clinical extraction tasks. Hardware as simple as an 8 GB laptop or a 12 GB consumer GPU has been used to run real systems, with response times of a few seconds and a 100% top-3 diagnostic accuracy achieved with curated local clinical cases in one offline diagnostic system.

While these results are a significant advancement towards the goal of accessible medical AI, the depth of evidence is still limited, recently collected, and geographically Localized. Only one of the eight studies included was specifically set in a sub-Saharan African context, none reported a completed prospective clinical trial, and the specific safety benchmark showed that, while these models looked ready to go and seemed adept and articulate, the majority of consumer-hardware-deployable models did not successfully complete a basic clinical safety test. Based on this evidence, three priorities emerge for future work: 1) Safety and reliability testing of the kind shown to be possible by Shlyakhta (2026) should become a standard companion to any feasibility study or any accuracy benchmark; 2) the early results from the Aletheia system in Uganda, demonstrated by Walusimbi et al. (2026), support further expansion of such systems into a larger prospective multi-site validation, which these authors call their next step; and 3) with the volume of this literature even in the 2024-2026 window covered in this piece, periodic updates to reviews such as this one will be needed to keep pace with a rapidly evolving field.

References

Al-Ganad, A., Al-Shahdhi, A., Al-Dhaifi, O., Hajeb, E., Hajeb, H., & Al-Motarreb, A. (2026). Deploying medical AI in low-resource settings: A scoping review of challenges and strategies. Frontiers in Digital Health, 8, 1743634. https://doi.org/10.3389/fdgth.2026.1743634

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arber, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

D'addario, A. M. V. (2025). Breaking the cost barrier: How quantization enables efficient development and deployment of LLMs for public healthcare. medRxiv. https://doi.org/10.1101/2025.11.17.25340460 (preprint, not peer-reviewed)

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314.

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2023). GPTQ: Accurate post-training quantization for generative pre-trained transformers. International Conference on Learning Representations (ICLR).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR).

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Kim, S. H., Schramm, S., Adams, L. C., Braren, R., Bressem, K. K., Keicher, M., Platzek, P.-S., Paprottka, K. J., Zimmer, C., Hedderich, D. M., & Wiestler, B. (2025). Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. npj Digital Medicine, 8, 97. https://doi.org/10.1038/s41746-025-01488-3

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., ... & Stoica, I. (2023). Efficient memory management for large language model serving with paged attention. Proceedings of the 29th Symposium on Operating Systems Principles (SOSP), 611-626.

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388(13), 1233-1239.

Lin, J., Tang, J., Tang, S., Chen, S., Yang, W., Chen, W. M., ... & Han, S. (2024). AWQ: Activation-aware weight quantization for LLM compression and acceleration. Proceedings of Machine Learning and Systems (MLSys), 6, 87-100.

Meskó, B., & Görög, M. (2020). A short guide for medical professionals in the era of artificial intelligence. NPJ Digital Medicine, 3(1), 126.

Meta AI. (2024). Llama 3 model card. GitHub. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259-265.

Rajbhandari, S., Rasley, J., Ruwase, O., & He, Y. (2020). ZeRO: Memory optimizations toward training trillion parameter models. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC20), 1-16.

Shakya, P. R., Khaneja, A., & Wagholikar, K. B. (2025). For clinical data extraction, QLoRA attains accuracy close to LoRA while requiring lower compute resources. PMC12633606. https://pmc.ncbi.nlm.nih.gov/articles/PMC12633606/

Shlyakhta, T. (2026). Benchmarking large language models for intensive care unit clinical decision support: A dual safety evaluation of 26 models on consumer hardware. medRxiv. https://doi.org/10.64898/2026.02.08.26345854 (preprint, not peer-reviewed)

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

Srinivasu, P. N., Samudrala, A., Devesh, P. G., Reddy, S. A., & Al Nuaim, A. (2026). Quantized low-rank adaptation in large language models for clinical text simplification. International Journal of Computational Intelligence Systems, 19, 267. https://doi.org/10.1007/s44196-026-01402-z

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., ... & Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Annals of Internal Medicine, 169(7), 467-473. https://doi.org/10.7326/M18-0850

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P. C., ... & Natarajan, V. (2024). Towards generalist biomedical AI. NEJM AI, 1(3), AIoa2300138.

Ugarte-Gil, C., Icochea, M., Llontop Otero, J. C., Villaizan, K., Young, N., Cao, Y., Liu, B., Griffin, T., & Brunette, M. J. (2020). Implementing a socio-technical system for computer-aided tuberculosis diagnosis in Peru: A field trial among health professionals in resource-constraint settings. Health Informatics Journal, 26(4), 2762-2775.

Valli, A. D., Tingle, S. J., Kazerouni, S., Kalpana, T. S. R., Karki, B., Knight, S. R., Wilson, C., & Kourounis, G. (2026). Artificial intelligence in surgical care within low-income and middle-income countries: A scoping review of development, validation, and deployment. eClinicalMedicine, 94, 103836. https://doi.org/10.1016/j.eclinm.2026.103836

Voinea, S.-V., Mămuleanu, M., Teică, R. V., Florescu, L. M., Selișteanu, D., & Gheonea, I. A. (2024). GPT-driven radiology report generation with fine-tuned Llama 3. Bioengineering, 11(10), 1043. https://doi.org/10.3390/bioengineering11101043

Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? BMJ Global Health, 3(4), e000798.

Walusimbi, J., Oguti, A. M., Sserwadda, A., Kasasira, P. B., & Okoboi, C. B. (2026). Aletheia: An offline-first clinical decision support system for differential diagnosis in low-resource healthcare settings. arXiv preprint arXiv:2607.24814.

World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2023). SmoothQuant: Accurate and efficient post-training quantization for large language models. Proceedings of the 40th International Conference on Machine Learning (ICML), 38087-38099.

Zhan, Z., Zhou, S., Zeng, M., Yu, K., Song, M., Chen, X., Wang, J., Hou, Y., & Zhang, R. (2025). Quantized large language models in biomedical natural language processing: Evaluation and recommendation. arXiv preprint arXiv:2509.04534.