UFSM

InterAção

ISSN 2357-7975
OPEN ACCESS

IA., Santa Maria, v. 17, n. 2, e97015, p. 1-20, 2026 • http://dx.doi.org/10.5902/2357797597015
Submitted: 1st/06/2026 • Approved: 31st/08/2026 • Published: 4th/09/2026

Artigos Dossiê

Performance of local Medical Large Language Models as clinical decision support systems under connectivity constraints

Desempenho de Large Language Models (LLMs) médicos locais como sistemas de suporte à decisão clínica, sob restrições de conectividade

Eduardo Borba Neves[I]
Pablo Gustavo Cogo Pochmann[I]

[I]Escola de Aperfeiçoamento de Oficiais, Rio de Janeiro, RJ, Brazil

Abstract

In the context of military health support, medical teams frequently operate under extreme conditions, characterized by severe connectivity restrictions, high cognitive pressure, the need for constant mobility, and severe limitations in computational and logistical resources. In this sense, the objective of this study was to investigate the performance of Large Language Models (LLMs), fine-tuned for medical applications, executed locally (offline), considering their application in operational scenarios with connectivity constraints. The experiment followed the classic validation pipeline of Clinical Decision Support Systems (CDSS). The experiment was executed using Python code on two hardware architectures: with a GPU and with a CPU (without a GPU). Eight models were tested using three reference medical benchmarks. The evaluation metrics were: Clinical Accuracy (%), Average Latency per Question (in seconds), and Tokens per Second (TPS). The results show that quantized models in the 8-billion parameter range, especially LlaMA-3.1-8B-Medical (Q4_K_M), have significant potential to act as force multipliers, achieving accuracies of up to 74% in complex medical benchmarks (such as MMLU_Medical and MedMCQA). However, artificial intelligence tools must be integrated into the battlefield exclusively as cognitive assistants under human supervision.

Keywords: Clinical decision support systems; Local Large Language Models, Military healthcare support; Offline operational environments; Model quantization

Resumo

No contexto do apoio de saúde militar, as equipes médicas frequentemente atuam sob condições extremas, caracterizadas por severas restrições de conectividade, elevada pressão cognitiva, necessidade de mobilidade constante e limitações severas de recursos computacionais e logísticos. Neste sentido, o objetivo deste estudo foi investigar o desempenho de modelos de Large Language Models (LLMs), fine-tuned para aplicações médicas, executados localmente (offline), considerando sua aplicação em cenários operacionais com restrições de conectividade. O experimento seguiu o pipeline clássico de validação de Sistemas de Suporte à Decisão Clínica (Clinical Decision Support Systems - CDSS). O experimento foi executado por meio de um código em Python, em duas estruturas de hardware: com GPU e com CPU (sem GPU). Foram testados oito modelos, utilizando três benchmarks médicos de referência. As métricas de avaliação foram: Acurácia Clínica (%), Latência Média por Questão (em segundos), e Tokens por Segundo (TPS). Os resultados evidenciam que modelos quantizados na faixa de 8 bilhões de parâmetros, em especial o LlaMA-3.1-8B-Medical (Q4_K_M), possuem um potencial significativo para atuar como multiplicadores de força. Ao alcançar acurácias de até 74% em benchmarks médicos complexos (como MMLU_Medical e MedMCQA). Contudo, as ferramentas de inteligência artificial devem ser integradas ao espaço de batalha exclusivamente como assistentes cognitivos sob a supervisão humana.

Palavras-chave: Sistemas de suporte à decisão clínica, Modelos de linguagem de grande porte locais, Apoio de saúde militar, Ambientes operacionais desconectados, Quantização de modelos

INTRODUCTION

The increasing complexity of contemporary military operations has expanded the need for robust technological solutions capable of supporting healthcare decision-making processes across diverse operational environments. Within the context of military medical support, medical teams frequently operate under extreme conditions characterized by severe connectivity constraints, high cognitive pressure, the need for constant mobility, and strict computational and logistical resource limitations. These conditions are formally established in defense doctrine as DIL (Delayed, Disconnected, Intermittently Connected, Low-Bandwidth) environments, where the network communication infrastructure is intermittent, degraded, or completely nonexistent (*Rivera-Nichols et al., 2025*). Consequently, traditional systems based on cloud computing or reliant on continuous data links prove inadequate and vulnerable for supporting timely medical assistance.

During high-intensity tactical operations, military healthcare professionals face the challenge of performing triage, stabilization, and therapeutic interventions under severe physical and psychological stress, routinely dealing with mass casualty scenarios (*Mathais et al., 2026*). Evidence in the military medicine literature indicates that operational environments of this nature significantly reduce the working memory capacity, selective attention, and critical decision-making efficacy of healthcare officers and enlisted personnel (*Angthong, Rungrattanawilai e Pundee, 2023*). Additionally, the persistent reliance on printed protocols and analog manuals at forward echelons of care imposes barriers to rapid and precise access to structured clinical information during patient care in combat environments (*Ravisankar, 2026*).

Within this framework of operational constraints, Large Language Models (LLMs) emerge as a promising technological alternative for real-time clinical support, enabling rapid consultation of field protocols, assistance with differential diagnosis, and general support for medical reasoning. Recent scientific investigations demonstrate that LLMs exhibit high proficiency in tasks such as medical record summarization, specialized biomedical information retrieval, the structuring of clinical management plans, and the development of conversational systems applied to healthcare (*Nazi e Peng, 2024*) (*Wang e Zhang, 2024*). However, the vast majority of these applications were architected with a direct dependence on centralized cloud infrastructures, which precludes their tactical employment in theaters of operations conditioned by DIL constraints.

The implementation of LLM models executed on a strictly local (offline) architecture offers crucial strategic and tactical advantages for the military ecosystem. Key benefits include immunity to external network unavailability, mitigation of inference latency, operational system resilience, and absolute control over the security, sovereignty, and confidentiality of the processed clinical and operational data. These factors assume critical relevance in the military domain, given that the transmission of biometric and geolocation data through open channels can compromise information security and jeopardize mission success (*Rivera-Nichols et al., 2025*). From a logistical

perspective, local inference confers substantial autonomy to medical teams deployed at forward aid stations, field hospitals, and isolated detachments operating in remote areas or under electronic warfare scenarios.

Despite the rapid progress of LLMs in the biomedical field, scientific literature highlights severe limitations associated with clinical accuracy, the occurrence of hallucinations (the generation of plausible but factually incorrect information), intrinsic biases in training data, and inconsistencies in the interpretation of highly complex clinical variables (*Nazi e Peng, 2024*), (*Wornow et al., 2023*), (*He et al., 2025*). The impact of these limitations is amplified in military environments, where decision-making deviations can lead to preventable fatalities and compromise the logistical readiness of the force. Concurrently, there is a methodological gap in studies focused on the empirical evaluation of LLMs operating in disconnected tactical scenarios under DIL constraints, particularly within the scope of Latin American armed forces and the operational specificities of the Brazilian Army.

Given the scenario described, this study investigates the performance of LLM models fine-tuned for medical applications and executed offline, considering their deployment in operational scenarios limited by DIL constraints. It sought to measure the potential and performance of such technologies as tools for tactical medical decision support, quantifying their efficacy and computational behavior under severe infrastructure constraints and prompt-response demands.

METHODOLOGY

This study utilized an experimental approach to evaluate the performance and computational efficiency of specialized medical domain LLMs subjected to distinct hardware infrastructure constraints. The methodological protocol was developed using Python and can be conceptually divided into four stages, as presented in Section 2.1.

Methodology Overview

The experiment follows the classical validation pipeline for Clinical Decision Support Systems (CDSS) in Clinical and Biomedical Engineering: structured data processing, local inference execution under rigorous variable control, and automated extraction of quantitative performance and computational efficiency metrics (*Xu et al., 2023*).

Stage 1 — Input Layer: Biomedical Databases

The pipeline is initialized by reading local Excel (.xlsx) files corresponding to three databases well-established in the biomedical Natural Language Processing (NLP) literature:

- MEDQA: Multiple-choice questions derived from the United States Medical Licensing Examination (USMLE), designed to assess integrative clinical reasoning and multi-step differential diagnosis (*Jin et al., 2021*).
- MedMCQA: A dataset containing over 180,000 questions from Indian medical entrance examinations (AIIMS/NEET), covering a wide diversity of medical specialties and theoretical concepts (*Pal, Umapathi e Sankarasubbu, 2022*).
- MMLU Medical Subjects: A subset of the Massive Multitask Language Understanding suite, filtered for clinical medicine, human biology, and genetics topics, constituting the global gold standard for evaluating encyclopedic knowledge in the biomedical sciences (*Hendrycks et al., 2020*).

To ensure strict reproducibility and execution feasibility, the pipeline applies a data cleaning routine—removing records with missing alternatives or undefined answer keys—followed by stratified random sampling (N = 100 questions per database). The consistency of the sampling across distinct runs is ensured by a fixed randomness seed (random_state = 42).

Stage 2 — Processing Layer: Local Inference Pipeline

The execution engine is unified for both hardware environments and comprises three sequential steps:

1. Model Loading via llama.cpp: The script identifies files with the .gguf extension in the model's directory and loads them individually into memory. All models employ 4-bit quantization (Q4), which drastically reduces the RAM/VRAM footprint without structural impact on the model architecture.
2. Prompt Engineering: The question and its four alternatives (A, B, C, D) are injected into a structured prompt template. This prompt defines the model's persona ("You are a specialist physician") and rigidly restricts the output behavior ("respond ONLY with the correct letter").
3. Inference Execution: The engine performs the inference configured with Temperature = 0.0, which eliminates text generation randomness, forcing the model to adopt a purely deterministic behavior (enhancing clinical reproducibility). The generation limit is locked at max_tokens = 10, and the newline character (\n) is defined as a stop token to avoid redundant text.

Stage 3 — Distribution Layer: Operational Computational Scenarios

The pipeline bifurcates to evaluate the system's behavior across two radically distinct hardware infrastructures, in accordance with the possibilities of military healthcare logistics.

Stage 4 — Output Layer: Metrics and Comparative Analysis

Once inference is complete, the raw results pass through a post-processing function based on regular expressions (re.search), which uniquely extracts the letter of the alternative selected by the model, normalizing outputs to the {A, B, C, D} pattern. Responses that do not correspond to any valid alternative are classified as "INVALID". The three metrics consolidated at the end of the pipeline are described in Section 2.5.

The results are automatically exported to a spreadsheet (.xlsx) including the date and time of execution.

Evaluated Models and Justifications

Eight open-source models in GGUF format, quantized in 4 to 8 bits and optimized for local execution via the llama.cpp library, were selected. The selection is based on three converging criteria: (i) deployment feasibility on consumer hardware without reliance on cloud infrastructure; (ii) existence of specialized fine-tuning within the biomedical domain; and (iii) coverage of different architectural families and parameter scales, enabling a comprehensive comparative analysis (*Touvron et al., 2023*). Table 1 summarizes the evaluated models.

Table 1 – Models evaluated in the experiment, identified by GGUF file, parameter count, quantization scheme, and base architecture

Model (GGUF file)	Parameters	Quantization	Architecture / Specialization
Bio-Medical-Llama-3.1-8B.i1-Q4_K_M	8B	Q4_K_M (i1)	Llama 3.1 — general biomedical fine-tuning
LlaMA-3.1-8B-Medical.Q4_K_M	8B	Q4_K_M	Llama 3.1 — clinical biomedical fine-tuning
llama-3-8b-chat-doctor.Q4_K_M	8B	Q4_K_M	Llama 3.0 — aligned for anamnesis dialogue
Llama-3.2-3B-Instruct-Medical-Conversational.Q8_0	3B	Q8_0	Llama 3.2 — conversational medical instruction, ultra-compact
Qwen2.5-3B-instruct-medical-finetuned.Q4_K_S	3B	Q4_K_S	Qwen 2.5 — medical fine-tuning, optimized for restricted hardware
Qwen2.5-3B-instruct-medical-finetuned-Heretic.i1-Q4_K_M	3B	Q4_K_M (i1)	Qwen 2.5 — Heretic variant with i1 quantization
Gemma-4-E2B-it-sft-rlvr-medical.Q4_K_S	~2B	Q4_K_S	Gemma 4 — medical RLHF/RLVR, advanced behavioral alignment
Medicine-llm.Q4_K_M	N/D	Q4_K_M	Proprietary general clinical medicine LLM

Source: Authors (2026)

Evaluation Databases (Benchmarks)

The selection of three distinct databases aims to mitigate memorization bias and evaluate heterogeneous medical competencies (differential diagnosis, anatomy, pharmacology, and encyclopedic clinical knowledge) in line with well-established practices in the biomedical Natural Language Processing (NLP) literature (*Jin et al., 2021*). For each database, a stratified random sampling of N = 100 questions was performed, with a fixed seed (random_state = 42) to ensure strict pipeline reproducibility across independent runs.

Hardware Scenarios and Operational Environment

The execution infrastructure was partitioned into two functional classes to map the behavior of the models under different military logistical constraints. Table 2 summarizes the specifications for each scenario.

Table 2 – Hardware specifications of the two evaluated operational scenarios

Component	Scenario 1 — Forward Medical Station (Limited Hardware)	Scenario 2 — Forward Medical Station (Medium-Capacity Hardware)
Processor	Intel Core i5-1235U (12th Gen, 10 cores, up to 4.40 GHz)	11th Gen Intel(R) Core(TM) i7-11700KF @ 3.60GHz (3.60 GHz)
RAM	12 GB DDR4	32 GB DDR4/DDR5
GPU	Intel Iris Xe (integrated, no dedicated VRAM)	Dedicated GPU with 8 GB VRAM (NVIDIA GeForce RTX 3070)
Storage	M.2 NVMe SSD	M.2 NVMe SSD
llama.cpp configuration	n_threads=8 (CPU execution)	n_gpu_layers=-1 (full offloading to VRAM)

Source: Authors (2026)

Scenario 1 — Forward Medical Station (Limited Hardware)

This scenario represents the typical infrastructure of mobile pre-hospital care (PHC) and field-degraded telemedicine. The Intel Core i5-1235U processor features a hybrid architecture with two Performance cores (P-cores) and eight Efficient cores

(E-cores). The absence of dedicated VRAM forces execution predominantly via the CPU, with parallelization configured at n_threads = 8, a value that symmetrically maps the E-core cluster without inducing thermal throttling—a phenomenon that would compromise the consistency of latency measurements.

Scenario 2 — Forward Medical Station (Medium-Capacity Hardware)

This represents the provision of medium-capacity hardware equipment to the forward medical station. The dedicated GPU with 8 GB of VRAM allows for full offloading of the model layers (n_gpu_layers = -1), transferring tensor multiplication computations to the high-bandwidth memory of the GPU and freeing the CPU for I/O operations and pipeline control.

Justification of Configuration Parameters

The context window was fixed at 4,096 tokens, a value substantially lower than the native windows of models such as Llama 3.1 (131,072 tokens) and Qwen 2.5, due to memory viability reasons. The KV Cache of a model with 32 layers, 8 KV heads, and a head dimension of 128, in FP16 precision, requires approximately 128 KB per token. Expanding the context to the native value would require around 16 GB exclusively for the KV cache, causing Out-Of-Memory (OOM) errors in both scenarios. The 4,096-token limit imposes a footprint of only an additional 512 MB, which is compatible with the hardware in Scenario 1, while simultaneously providing an ample safety margin for multiple-choice questions, whose text rarely exceeds 500 to 800 tokens.

The number of threads = 8 (eight) was established based on the law of diminishing returns of parallel computing. Hyperthreading technology (SMT) duplicates the logical cores exposed to the operating system but leaves the physical Floating-Point Unit (FPU) shared. Since neural network inference is computationally intensive regarding floating-point calculations, saturating the FPU with two concurrent threads induces context switching and degrades the Tokens per Second (TPS) rate. Configuring one thread per actual physical core guarantees the locality and coherence of L2 and L3 cache memories, maximizing throughput efficiency in both scenarios (Dongarra *et al*., 2024).

Performance and Efficiency Metrics

The pipeline synchronously collects three fundamental metrics to determine the clinical and operational viability of AI systems (Touvron *et al*., 2023) (Dongarra *et al*., 2024) (Nguyen *et al*., 2025):

- **Clinical Accuracy (%):** Maps the accuracy rate by comparing the predicted response with the actual answer key of the question (correct_answer).
- **Average Latency per Question (s):** Measures the average time in seconds (calculated via the time library) that the hardware took to process the prompt and complete the response generation.
- **Tokens per Second (TPS):** Measures computational throughput efficiency by dividing the number of generated tokens by the time spent in the process.

RESULTS

Experimental Configuration

Eight Large Language Models (LLMs) subjected to medical fine-tuning were evaluated, all executed in GGUF format via the llama.cpp framework. The selected models represent different architectural families—Llama 3.1 (8B parameters), Llama 3.2 (3B), Qwen 2.5 (3B), Gemma 4 (2B), and Medicine-LLM (7B)—with quantization schemes varying between Q4_K_S, Q4_K_M, and Q8_0, allowing for the evaluation of compression impact on response quality.

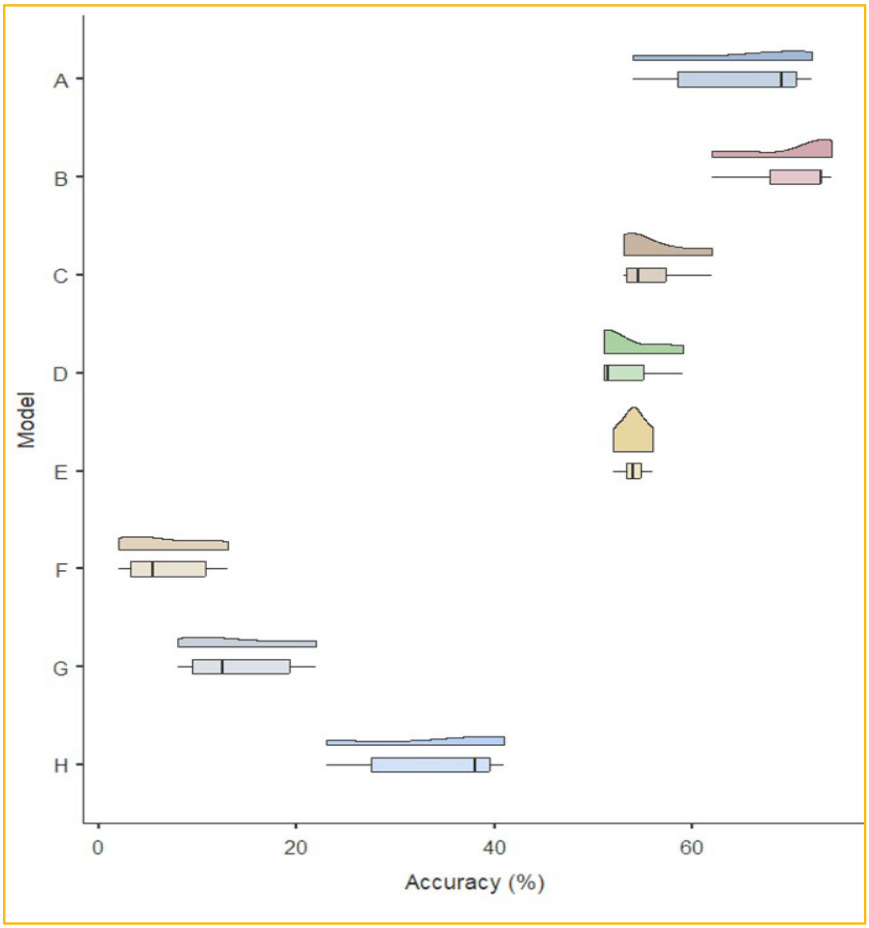
The experiments were conducted across two hardware configurations representative of the military operational context in communication-constrained environments (DIL): (i) CPU-only with 12 GB of RAM, simulating lightweight, low-cost devices; and (ii) a GPU with 8 GB of VRAM combined with 32 GB of RAM, representing portable workstations with graphics acceleration.

Performance was measured using three reference medical benchmarks: **MEDQA**

(multiple-choice questions from medical licensing examinations, with 4 options per question); **MMLU_Medical** (the medical sciences subset of the Massive Multitask Language Understanding suite); and **MedMCQA** (a question bank from medical entrance examinations, with 4 alternatives). For each model–dataset–hardware combination, accuracy (%), average latency per question (seconds), and token generation rate per second (TPS) were recorded.

Figure 1 displays the accuracy obtained by the models considering both hardware variations. Detailed results for accuracy, latency, and TPS will be presented in Sections 3.2 and 3.3.

Figure 1 – Accuracy obtained by the models across the three reference medical benchmarks: MEDQA, MMLU_Medical, and MedMCQA



Source: Authors (2026)

Legend: A: Bio-Medical-Llama-3.1-8B (Q4_K_M), B: LlaMA-3.1-8B-Medical (Q4_K_M), C: Llama-3.2-3B-Med-Conv (Q8_0), D: Qwen2.5-3B-med-Heretic (Q4_K_M), E: Qwen2.5-3B-med-finetuned (Q4_K_S), F: gemma-4-E2B-medical (Q4_K_S), G: llama-3-8b-chat-doctor (Q4_K_M), H: medicine-llm (Q4_K_M).

Performance on CPU-Only Hardware (12 GB RAM)

Table 3 presents the results obtained in the CPU-only configuration with 12 GB of RAM. In this scenario, which simulates low-cost and highly portable devices, the models from the Llama 3.1 family with 8 billion parameters demonstrated consistent superiority in accuracy.

Table 3 – Performance results in the CPU configuration (12 GB RAM). Acc. = Accuracy; Lat. = Average latency; TPS = Tokens per second

Model	MEDQA			MMLU_Medical			MedMCQA		
	Ac. (%)	Lat. (s)	TPS	Ac. (%)	Lat. (s)	TPS	Ac. (%)	Lat. (s)	TPS
Bio-Medical-Llama-3.1-8B (Q4_K_M)	54	10,27	0,87	69	8,23	1,07	72	3,66	2,16
LlaMA-3.1-8B-Medical (Q4_K_M)	66	10,44	0,96	73	8,30	1,14	73	3,69	2,18
Llama-3.2-3B-Med-Conv (Q8_0)	58	7,33	1,22	54	5,93	1,46	53	2,60	3,05
Qwen2.5-3B-med-Heretic (Q4_K_M)	51	4,18	1,79	56	3,40	1,91	51	1,50	3,93
Qwen2.5-3B-med-finetuned (Q4_K_S)	53	3,72	1,47	56	3,04	1,63	54	1,21	3,47
Gemma-4-E2B-medical (Q4_K_S)	7	2,35	0,48	3	1,85	0,59	13	0,67	1,67
Llama-3-8b-chat-doctor (Q4_K_M)	14	8,83	0,23	21	7,15	0,28	9	2,40	0,57
Medicine-llm (Q4_K_M)	24	12,19	0,90	40	9,99	1,08	38	4,55	2,29

Source: Authors (2026)

The LlaMA-3.1-8B-Medical (Q4_K_M) model achieved the best absolute results in two of the three benchmarks: 66% accuracy on MEDQA and 73% on both MMLU_ Medical and MedMCQA. Bio-Medical-Llama-3.1-8B (Q4_K_M) presented competitive performance, with accuracies of 54%, 69%, and 72% across the same benchmarks, respectively. Both models operate with high latencies in the CPU configuration, ranging between 3.66 and 10.44 seconds per question, and generation rates between 0.87 and 2.18 TPS, reflecting the lack of acceleration by dedicated hardware.

The Llama-3.2-3B-Instruct-Medical-Conversational (Q8_0) model, featuring 3 billion parameters with Q8_0 quantization, exhibited lower accuracies of 58%, 54%, and 53%, but with lower latencies (2.60–7.33 s) and higher TPS (up to 3.05), making it an intermediate option when response speed is prioritized over accuracy.

The models from the Qwen2.5-3B family (in the Heretic Q4_K_M and finetuned Q4_K_S variants) displayed modest accuracies (51–56%) with significantly lower latencies (1.21–4.18 s) and higher TPS (up to 3.93), making them the fastest models in the CPU configuration. However, their accuracy may be insufficient for critical clinical decisions.

The gemma-4-E2B-medical (Q4_K_S), llama-3-8b-chat-doctor (Q4_K_M), and medicine-llm (Q4_K_M) models showed unsatisfactory performance across all benchmarks. Gemma 4E2B recorded accuracy of only 7%, 3%, and 13%, highlighting its inadequacy for medical use in the evaluated configuration. Despite its 7-billion-parameter architecture, medicine-llm did not exceed 40% accuracy on any dataset, with latencies up to 12.19 s, the highest in the entire experiment.

Performance on GPU (8 GB VRAM) + 32 GB RAM Hardware

Table 4 displays the results obtained with the GPU + extended RAM configuration. VRAM acceleration significantly reduced latencies across all models, with a more pronounced impact on larger models. The accuracy hierarchy among the models remained essentially unchanged compared to the CPU configuration.

The LlaMA-3.1-8B-Medical (Q4_K_M) remained the model with the highest overall accuracy, reaching 62%, 73%, and 74% on MEDQA, MMLU_Medical, and MedMCQA, respectively. With the GPU, its average latency dropped to the 2.14–5.02 s range and its TPS reached 3.67, representing a speed improvement of approximately 60–70% compared to the CPU configuration.

The Bio-Medical-Llama-3.1-8B (Q4_K_M) presented similar performance to the former, with accuracies of 55%, 69%, and 71%, and latencies of 2.07–4.85 s. The

difference between the two Llama 3.1 8B models is marginal in accuracy, but LlaMA-3.1-8B-Medical showed a consistent, slight advantage across all three benchmarks.

Table 4 – Performance results in the GPU (8 GB VRAM) + 32 GB RAM configuration. Acc. = Accuracy; Lat. = Average latency; TPS = Tokens per second

Model	MEDQA			MMLU_Medical			MedMCQA		
	Ac. (%)	Lat. (s)	TPS	Ac. (%)	Lat. (s)	TPS	Ac. (%)	Lat. (s)	TPS
Bio-Medical-Llama-3.1-8B (Q4_K_M)	55	4,85	1,79	69	4,11	2,10	71	2,07	3,77
LlaMA-3.1-8B-Medical (Q4_K_M)	62	5,02	1,98	73	4,20	2,22	74	2,14	3,67
Llama-3.2-3B-Med-Conv (Q8_0)	62	4,41	1,99	55	3,73	2,28	53	1,87	4,22
Qwen2.5-3B-med-Heretic (Q4_K_M)	52	2,10	3,66	59	1,74	4,05	51	0,83	7,23
Qwen2.5-3B-med-finetuned (Q4_K_S)	54	1,83	2,95	55	1,44	3,30	52	0,67	6,16
Gemma-4-E2B-medical (Q4_K_S)	4	1,12	1,00	2	0,88	1,25	12	0,33	3,32
Llama-3-8b-chat-doctor (Q4_K_M)	11	3,79	0,53	22	3,01	0,66	8	1,10	1,23
Medicine-llm (Q4_K_M)	23	6,39	1,69	41	5,18	2,05	38	2,86	3,58

Source: Authors (2026)

The Llama-3.2-3B-Medical-Conversational (Q8_0) notably benefited from the GPU: on MEDQA, its accuracy rose from 58% (CPU) to 62% (GPU), matching the performance of the 8B models on this benchmark. However, on MMLU_Medical and MedMCQA, its accuracy remained lower (55% and 53%), suggesting that the gain is more pronounced in direct clinical questions rather than broad medical reasoning.

The models from the Qwen2.5-3B family demonstrated the highest relative gain in TPS with the GPU; the Heretic variant reached 7.23 TPS on MedMCQA, the highest value in the entire experiment, though without a proportional improvement in accuracy (51–59%). For applications that prioritize response speed over diagnostic precision, these models offer a relevant operational advantage.

The gemma-4-E2B, llama-3-8b-chat-doctor, and medicine-llm models did not show significant accuracy improvements with the GPU, reinforcing that their limitations are intrinsic to the applied fine-tuning rather than related to the inference hardware.

DISCUSSIONS

Experimental results demonstrate the viability and challenges of implementing LLMs as CDSS in DIL environments (*Rivera-Nichols et al., 2025*). In the context of 21st-century combat casualty care, tactical environments frequently impose severe limitations on resources, communication infrastructure, and access to specialists (*Ravisankar, 2026*). The use of digital tools and offline artificial intelligence capable of providing autonomous support is highlighted as a necessary evolution to modernize the battlespace, replacing paper protocols and assisting responders under high cognitive load (*Rivera-Nichols et al., 2025*) (*Leone et al., 2024*) (*Riggenbach et al., 2025*).

The superior performance of the LlaMA-3.1-8B-Medical (Q4_K_M) model, which achieved accuracies of up to 73% and 74% on the MMLU_Medical and MedMCQA benchmarks, confirms the ability of 8-billion-parameter models to retain and apply professional-level medical knowledge. The MMLU tests language understanding across multiple tasks, encompassing knowledge and problem-solving that require an extensive intellectual background (*Hendrycks et al., 2020*), whereas MEDQA demands complex logical reasoning steps regarding symptoms and patient presentations to arrive at a diagnosis (*Jin et al., 2021*). Achieving consistent performance at these levels while operating 100% offline meets a critical demand in military medicine: the need for reliable prolonged field care support, where life-saving interventions are heavily influenced by the level of available expertise (*Ravisankar, 2026*). However, high benchmark accuracy should not be the sole metric of success; it is imperative to recognize that standardized evaluations may fail to capture the true clinical utility and operational bottlenecks of the technology in practice (*Wornow et al., 2023*).

The evaluation across two hardware profiles highlighted the inherent speed-accuracy trade-off, corroborating with recent literature showing that accuracy gains in LLMs geared toward medical reasoning come at the cost of substantial latency penalties (*Nguyen et al., 2025*). In the CPU-only scenario (simulating highly portable devices), the latencies of Llama 3.1 8B exceeded 10 seconds per question. In tactical environments requiring split-second triage responses or hemorrhage control, response delays can be critical in the management of polytrauma (*Angthong, Rungrattanawilai e Pundee, 2023*). The use of quantization (such as Q4 and Q8) mitigates part of this problem, reflecting an essential hardware trend to balance numerical precision, power consumption, and computational capacity in resource-constrained systems (*Dongarra et al., 2024*). When accelerated by a GPU, the 8B model became significantly smoother (latency reduced to ~2 to 5 seconds), indicating that advanced or command workstations are the ideal platforms to support deep diagnostic reasoning.

For scenarios where response speed is the absolute priority, smaller models such as Llama-3.2-3B-Med-Conv and variants of the Qwen2.5-3B family exhibited substantially higher TPS and lower latency. This computational agility is valuable for military algorithm-based tactical triage operations, dynamically assisting in treatment decisions for shock or basic life support. However, their lower accuracy (in the 51% to 62% range) confines them to subsidiary and low-clinical-risk tasks, where the model acts more as a protocol extraction assistant rather than an independent diagnostic engine (*Nguyen et al., 2025*). It is well-documented that models optimized for speed fail more frequently on questions requiring complex multi-step reasoning, thus necessitating robust human supervision (*Nguyen et al., 2025*).

The unsatisfactory performance of models such as gemma-4-E2B-medical, llama-3-8b-chat-doctor, and medicine-LLM highlights the critical challenges of the fine-tuning process for the medical domain. Literature indicates that training LLMs for healthcare requires highly curated, quality data and a robust alignment methodology (*Nazi e Peng, 2024*) (*Wang e Zhang, 2024*) (*Wornow et al., 2023*) (*He et al., 2025*). Models

trained primarily on conversational dialogues or on small/unbalanced medical datasets frequently fail to generalize structured and highly technically demanding clinical tasks (*Wang e Zhang, 2024*). Furthermore, LLMs without proper alignment carry the risk of producing "hallucinations" and fabricated information, an unacceptable hazard in a medical decision-making context, where the slightest error can result in fatalities (*Nazi e Peng, 2024*) (*Wang e Zhang, 2024*).

Finally, the use of LLMs in military and healthcare settings requires considerations beyond technical performance. The interpretability and transparency of model recommendations (white box vs. black box) are fundamental for military healthcare professionals to trust CDSS guidance under stress (*Xu et al., 2023*) (*Hoyt et al., 2025*). None of the tested models reached a level of clinical autonomy capable of replacing the judgment of a responder or physician; instead, they act as force multipliers (*Riggenbach et al., 2025*) Therefore, their operational adoption on the battlefield must be linked to clear guidelines, expert supervision, and addressing ethical limitations regarding civil liability and the inherent biases in the training of these networks (*Wang e Zhang, 2024*) (*Wornow et al., 2023*) (*He et al., 2025*).

CONCLUSIONS

This study achieved its objective by demonstrating the viability and quantifying the performance of Large Language Models (LLMs) fine-tuned for the medical field, executed entirely locally (offline) as clinical decision support tools in tactical operational environments constrained by DIL (Delayed, Disconnected, Intermittently Connected, Low-Bandwidth) limitations.

The results demonstrate that quantized models in the 8-billion-parameter range, particularly LlaMA-3.1-8B-Medical (Q4_K_M), possess significant potential to act as force multipliers. By achieving accuracies of up to 74% on complex medical benchmarks (such as MMLU_Medical and MedMCQA), the technology proves its efficacy in retaining and applying specialized, professional-level medical knowledge, bridging a

critical gap in access to specialists on the battlefield. Conversely, for scenarios where response speed is the absolute priority, smaller models such as Llama-3.2-3B-Med-Conv and variants of the Qwen2.5-3B family exhibited substantially higher TPS and lower latency. This computational agility is valuable for algorithm-based tactical triage military operations, dynamically supporting treatment decisions for shock or basic life support.

In conclusion, the application of LLMs in tactical military medicine is highly promising, but the technology does not yet possess the transparency (interpretability) or maturity to operate autonomously. The evaluated AI tools should be integrated into the 21st-century battlespace exclusively as cognitive assistants under human supervision. The effective modernization of military pre-hospital care through these technologies will depend on the provision of balanced hardware (portable GPUs), rigorously trained models, and the formulation of ethical and operational guidelines that always place the human responder at the center of decision-making.

REFERENCES

ANGTHONG, C.; RUNGRATTANAWILAI, N.; PUNDEE, C. Artificial intelligence assistance in deciding management strategies for polytrauma and trauma patients. **Polish Journal of Surgery**, 96, n. Suppl. 1, 2023. 114-117.

DONGARRA, J. *et al.* Hardware trends impacting floating-point computations in scientific applications. **arXiv preprint arXiv:2411.12090**, 2024.

HE, K. *et al.* A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. **Information Fusion**, 118, 2025. 102963.

HENDRYCKS, D. *et al.* Measuring Massive Multitask Language Understanding. **arXiv preprint arXiv:2009.03300**, 2020.

HOYT, R. E. *et al.* Evaluating Large Reasoning Model Performance on Complex Medical Scenarios In The MMLU-Pro Benchmark. **medRxiv preprint**, 2025. 2025-04.

JIN, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. **Applied Sciences**, 11, n. 14, 2021. 6421.

LEONE, R. M. *et al.* Artificial intelligence in military medicine. **Military Medicine**, 189, n. 9-10, 2024. 244-248.

MATHAIS, Q. *et al.* Artificial intelligence for battlefield triage in large-scale combat operations: opportunities, limits, and ethical considerations. **Journal of Trauma and Acute Care Surgery**, 100, n. 3, 2026. 412-420.

NAZI, Z. A.; PENG, W. Large Language Models in Healthcare and Medical Domain: A Review. **Informatics**, 11, n. 3, 2024. 57.

NGUYEN, V. A. *et al.* Quantifying the speed-accuracy trade-off of large language models on oral and maxillofacial surgery multiple-choice questions. **Scientific Reports**, 15, n. 1, 2025. 40657.

PAL, A.; UMAPATHI, L. K.; SANKARASUBBU, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. **Proceedings of the Conference on Health, Inference, and Learning**, 174, 2022. 248-260.

RAVISANKAR, A. V. Combat Casualty Care in the 21st Century: Advances, Challenges, and Evidence-Based Strategies for Armed Forces Medical Services. **JSES International**, 10, n. 101716, 2026.

RIGGENBACH, Z. W. *et al.* AI on the Front Lines: A Primer for the Military Health Professional. **Military Medicine**, 190, n. 9-10, 2025. e1851-e1857.

RIVERA-NICHOLS, T. *et al.* Investigation and Analysis of Available Chatbot Technologies to Integrate in Multi-Domain Operational, Delayed/Disconnected, Intermittently Connected, Low-Bandwidth Conditions. **Military Medicine**, 190, n. suppl. 2, 2025. 829-836.

TOUVRON, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023.

WANG, D.; ZHANG, S. Large language models in medical and healthcare fields: applications, advances, and challenges. **Artificial Intelligence Review**, 57, n. 11, 2024. 299.

WORNOW, M. *et al.* The shaky foundations of large language models and foundation models for electronic health records. **npj Digital Medicine**, 6, n. 1, 2023. 135.

XU, Q. *et al.* Interpretability of Clinical Decision Support Systems Based on Artificial Intelligence from Technological and Medical Perspective: A Systematic Review. **Journal of Healthcare Engineering**, 2023, 2023. 1-13.

Authorship

1 Eduardo Borba Neves

Doctor of Biomedical Engineering from the Federal University of Rio de Janeiro; Doctor of Public Health and Environment from the National School of Public Health – ENSP of the Oswaldo Cruz Foundation; Doctor of Notorious Knowledge in Military Education and Culture from the Department of Teaching and Research – DEP of the Brazilian Army; Coordinator

https://orcid.org/0000-0003-4507-6562 • neveseb@gmail.com

2 Pablo Gustavo Cogo Pochmann

Master of Military Sciences from the Officers' Improvement School – EsAO of the Brazilian Army; Master of Defense Engineering from the Military Engineering Institute – IME of the Brazilian Army; Professor

https://orcid.org/0000-0003-3944-7953 • pablo.pochmann@gmail.com

How to quote this article

NEVES, E. B.; POCHMANN, P. G. C. Performance of local Medical Large Language Models as clinical decision support systems under connectivity constraints. **InterAção**, Santa Maria, v. 17, n. 2, e97015, p. 1-20, 2026. DOI 10.5902/1980509897015. Available from: https://dx.doi.org/10.5902/2357797597015. Accessed in: day month abbr. year.