Letter to the Editor

Human-in-the-Loop as a Safety Guardrail: Clinical Accountability in the Large Language Model Era

Isaac Zablah[1*], PhD; Yolly Molina[2*], MSc; Antonio Garcia-Loureiro[3*], PhD

[1]Faculty of Medical Sciences, National Autonomous University of Honduras, Tegucigalpa, Honduras
[2]Center for Biomedical Imaging Diagnostics Research and Rehabilitation, National Autonomous University of Honduras, Tegucigalpa, Honduras
[3]Department of Electronics and Computer Science, Universidade de Santiago de Compostela, Santiago de Compostela, Spain
*all authors contributed equally

Corresponding Author:

Isaac Zablah, PhD
Faculty of Medical Sciences
National Autonomous University of Honduras
Calle la Salud SN
Tegucigalpa 11101
Honduras
Email: jose.zablah@unah.edu.hn

Related Article:
Comment on: https://www.jmir.org/2025/1/e59069

J Med Internet Res 2026;28:e85726; doi: 10.2196/85726

Keywords: large language models; high-performance computing; medical informatics; computational efficiency; clinical decision support; artificial intelligence; healthcare infrastructure; model optimization

We found Zhang et al's thorough review of the transformative potential of large language models (LLMs) in healthcare to be very interesting [1]. The authors do a great job of talking about clinical applications, data integration, and ethical issues. However, we think that important aspects of computational performance need more attention, especially when it comes to using technology in real-world healthcare settings where resources are limited.

Zhang et al mention that "*technological advancements*" are helping to meet the "*high hardware requirements*" of LLMs [1], but the reality of computing is still a huge challenge. Modern medical LLMs such as GPT-4 and domain-specific models such as Med-PaLM 2 need considerable infrastructure as described below [2]:

- *Inference latency*: Currently, LLMs take 2 to 10 seconds to respond to each query. This may not be fast enough for clinical situations where time is of the essence, like triage in the emergency department or decision support during surgery. More detailed answers need more time [3].
- *Memory footprint*: Models with billions of parameters need 16-80+ GB of VRAM (video random access memory) for fast inference [4]. This means that many health care facilities, especially in low- and middle-income countries, do not have the specialized GPU infrastructure they need.
- *Scalability challenges:* Serving hundreds of concurrent clinical users requires distributed computing architectures and load-balancing strategies not discussed in the review [5].

For edge computing and improving models, we suggest that subsequent research should emphasize:

- *Model quantization and pruning:* Techniques to reduce model size by 50%-75% with minimal accuracy loss, enabling deployment on consumer-grade hardware.
- *Edge computing solutions:* Local deployment using optimized models (eg, 7-13B parameter variants) to address data privacy concerns while reducing latency and cloud dependency.
- *Hybrid architecture:* Combining lightweight edge models for routine queries with cloud-based full models for complex cases, optimizing the accuracy-efficiency trade-off.

The medical informatics community requires standardized metrics that assess not only diagnostic accuracy but also operations per diagnosis (computational cost), energy consumption per inference (environmental impact), and cost-effectiveness ratios (accuracy gained per dollar of infrastructure). We did an initial benchmarking of three LLMs on differential diagnosis tasks: Clinical Camel (LLaMA-2-13B), PMC-LLaMA 13B, and Meditron-3 (Qwen2.5-14B). We found that smaller, domain-specific

models (~14 billion parameters fine-tuned on medical corpora) were able to achieve 85%-90% of GPT-4's diagnostic accuracy while using only about 15% of the computational resources, indicating considerable room for improvement.

We want high-performance computing research in medical artificial intelligence (AI) to help with clinical implementation. This research should set benchmarks for both computational performance and clinical accuracy, come up with optimization techniques that are specific to medical inference workloads, create reference architectures for deploying LLMs in different health care settings, and investigate federated learning strategies that let training happen without putting sensitive patient data in one place.

The transformative potential Zhang et al describe will only be realized if LLMs can be deployed efficiently and equitably across diverse health care environments. High-performance computing and medical informatics must advance in tandem to bridge the gap between research promise and clinical reality.

Acknowledgments

The authors used the Wordvice.ai service solely to improve the language and semantics of the manuscript.

Funding

The authors declared no financial support was received for this work.

Data Availability

The benchmarking data comparing diagnostic accuracy and computational resource utilization of Clinical Camel (LLaMA-2-13B), PMC-LLaMA 13B, and Meditron-3 (Qwen2.5-14B) against GPT-4 baseline are available from the corresponding author upon reasonable request. The evaluation was conducted on publicly available differential diagnosis case datasets. Model access: Clinical Camel and PMC-LLaMA 13B are available via Hugging Face; Meditron-3 (Qwen2.5-14B) is available through the EPFL repository; and GPT-4 was accessed via OpenAI API for comparative benchmarking.

Authors' Contributions

Conceptualization: AGL
Methodology: JZ
Validation: YM
Formal analysis: AGL
Writing — original draft: JZ, YM
Writing — review & editing: AGL

Conflict of Interest:

None declared.

Editorial Notice

The corresponding author of "Revolutionizing Health Care: The Transformative Impact of Large Language Models in Medicine" declined to respond to this letter.

References

1. Zhang K, Meng X, Yan X, et al. Revolutionizing health care: the transformative impact of large language models in medicine. J Med Internet Res. Jan 7, 2025;27:e59069. [doi: 10.2196/59069] [Medline: 39773666]
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature New Biol. Aug 2023;620(7972):172-180. [doi: 10.1038/s41586-023-06291-2] [Medline: 37438534]
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. Aug 2023;29(8):1930-1940. [doi: 10.1038/s41591-023-02448-8] [Medline: 37460753]
4. Raiaan MAK, Mukta MdSH, Fatema K, et al. A review on large language models: architectures, applications, taxonomies, open issues and challenges. IEEE Access. 2024;12:26839-26874. [doi: 10.1109/ACCESS.2024.3365742]
5. He K, Mao R, Lin Q, et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. arXiv. Preprint posted online on 2023. [doi: 10.48550/arXiv.2310.05694]

Abbreviations

AI: artificial intelligence
LLM: large language model
VRAM: video random access memory

Edited by Amaryllis Mavragani; This is a non–peer-reviewed article; submitted 12.Oct.2025; accepted 08.May.2026; published 18.Jun.2026

Please cite as:
Zablah I, Molina Y, Garcia-Loureiro A
Human-in-the-Loop as a Safety Guardrail: Clinical Accountability in the Large Language Model Era
J Med Internet Res 2026;28:e85726
URL: https://www.jmir.org/2026/1/e85726
doi: 10.2196/85726