Received: 14 OCT 2025 | Accepted: 20 OCT 2025 | **DOI:** https://doi.org/ 10.57849/ulisboa.fm.jscml.0000024.2025

JORNAL DA SOCIEDADE DAS CIÊNCIAS MÉDICAS DE LISBOA

VIEWPOINT

Large Language Models in Medicine

António Vaz Carneiro [1]

[1] Institute for Evidence Based Healthcare, Lisbon Academic Medical Center, Cochrane Portugal
✉ *avc@isbe.research.ulisboa.pt*

KEYWORDS: Large Language Models; Artificial Intelligence; Clinical Applications; Ethics and Regulation; Medical Education

Introduction

The application of artificial intelligence (AI) to medicine is not new, but the advancement of Large Language Models [LLMs], such as GPT-5 and other emerging multimodal models, represents a qualitative forward shift. These models have already moved beyond the experimental domain and are beginning to integrate into everyday clinical tasks, from supporting the writing of notes and reports, to triaging patient messages and synthesizing scientific literature [1,2].

The apparent benefits are important: increased efficiency, reduced administrative overhead, rapid access to information, interviewing patients and even support for diagnostic reasoning. However, the transition from enthusiasm to responsible adoption requires a new level of rigor: less technological fascination and more scientific maturity. In practice, the question has shifted from "can the LLM do this?" to "should it do it, and under what conditions?"

In this article we discuss some aspects of the role of LLMs in medicine. The potential applications of this technology are immense, but for this paper we selected some of the most relevant and well studied. We have included a glossary at the end, with some useful definitions.

Where LLMs already demonstrate value

Clinical Documentation and Summarization: Several studies have shown that models adapted to the medical context can outperform human experts in writing and summarizing clinical notes when supervised [3]. There is evidence that LLM trained with real clinical data produced summaries superior to those of physicians in terms of accuracy and completeness, saving significant administrative time [1]. These results point to a concrete opportunity: reducing bureaucratic burden and returning time to direct care.

Patient Communication: Models like ChatGPT have been tested in the context of responding to patient messages. In a study from JAMA Internal Medicine (2023), AI-generated responses were rated by an independent panel as more empathetic and detailed than those of physicians [4]. However, the same study emphasizes that AI should not replace the practitioner, but rather serve as a co-pilot, preparing drafts that the clinician reviews and validates.

Medical Education and Health Literacy: LLMs have demonstrated remarkable performance on standardized USMLE exams and other medical tests [5,6]. While these results demonstrate high performance and knowledge, they fail to capture contextual clini-

cal reasoning. In practice, the LLM can be useful as an educational tool (synthesizing information, explaining concepts, and generating clinical cases) but it still does not replace human clinical judgment.

Limitations and Risks

Despite advances, LLMs have serious limitations that cannot be ignored.

Hallucinations and Factual Errors: The phenomenon of hallucination (the generation of incorrect but plausible information) is perhaps the greatest clinical risk. Even recent models continue to produce false statements with high confidence, which can induce clinical error if the user blindly trusts the output [7].

Bias and Fairness: LLMs reproduce (and sometimes amplify) the biases present in the data they were trained on. This can translate into disparities in response quality by ethnic group, gender, language, or socioeconomic status [8]. Stratified performance testing and continuous adjustments are essential to ensure fairness.

Lack of Transparency and Traceability: The most powerful language models are often closed-source, making thorough audits impossible and error correction difficult. This opacity is incompatible with the principle of traceability that governs medical technologies. The scientific community has called for the adoption of open-source, auditable models for clinical use, with detailed documentation ("model cards") [9].

Transferability and Context: The performance of a model trained in one language or in one healthcare system can degrade substantially in other contexts. Local adaptation, via fine-tuning or contextual prompt engineering, is necessary to avoid cultural and clinical errors [10].

Ethics, Safety, and Regulation

In 2024, the World Health Organization [WHO] published the first ethical and governance guide for large-scale [multi]modal models in healthcare, with more than 40 recommendations [11]. These include: Independent prior evaluation of performance and safety; Post-deployment monitoring; Transparency of training data and limitations; Mandatory human oversight in critical decisions; Protection of personal data and informed consent.

In the US, the FDA updated its guidelines for software with AI/ML functions, integrating LLMs into the Software as a Medical Device [SaMD] category [12]. At the European level, the AI Act (2024) classifies medical AI as "high risk," requiring technical documentation, lifecycle management, explainability, and CE certification before commercialization [13].

Concurrently, the debate over legal liability is growing. If a clinical decision results from an interaction between a human and AI, who is liable in the event of harm? Recent literature advocates for clear institutional contracts and automatic records of AI-assisted decisions, ensuring shared accountability [14].

From Algorithmic Evidence to Clinical Evidence

Research on LLMs has been dominated by algorithmic metrics: accuracy, BLEU scores, F1 scores, pass rates. However, clinical utility requires another type of proof: real impact on health outcomes, efficiency, and safety. Pragmatic studies in hospital settings are still scarce. In 2024, the NEJM AI emphasized that "transition to clinical practice requires demonstrating direct benefit to patients, not just computational performance" [15].

Furthermore, equity assessment should become mandatory. LLM performance should be tested by age, gender, language, and socioeconomic status, and published, just as demographic composition is published in clinical trials. In the Table I we suggest operational recommendations for responsible implementation of LLMs in clinical settings.

The Near Future: Multimodality and Integration

The next step for LLMs in medicine is multimodality: models capable of processing text, images, audio, and structured data simultaneously. This will enable joint interpretations of reports, exams, and clinical examinations. One should be alert that each new type of data introduces new privacy risks and validation complexities [16].

Simultaneously, LLMs are being integrated into electronic health records [EHRs], facilitating information querying, automated note-writing, and interdisciplinary communication. These applications promise

TABLE I. Operational recommendations for responsible implementation of LLMs in clinical settings.

FASE	RECOMMENDED ACTION	OBJECTIVE
1 Usage selection	Choose low-risk/high-volume tasks (e.g., drafting notes, administrative responses)	Minimize risk and maximize impact
2 Prior assessment	DPIA (personal data impact assessment, legal opinion, and ethical analysis)	Legal and ethical compliance
3 Local adaptation	Light fine-tuning with anonymized data or contextual prompt engineering	Clinical and linguistic relevance
4 Independent validation	Testing with clinically meaningful metrics and diverse samples	Reliability and fairness
5 Governance	Documentation ("model cards"), version logs, rollback mechanisms	Transparency and traceability
6 Human oversight	Define the role of the human-in-the-loop per task	Clinical control and safety
7 Continuous monitoring	Periodic audits, performance logs, feedback channels	Continuous improvement
8 User training	AI literacy and training to identify hallucinations and bias	Safe use conscious

substantial efficiency gains but require improved control and explainability mechanisms [17].

Conclusion

Large language models [LLMs] are rapidly transforming medical practice, from administrative tasks to assisted clinical reasoning.

The enthusiasm surrounding these tools is understandable, but their potential benefits will only be realized if they are treated as medical technologies, with appropriate validation, oversight, and regulation. Large language models are not "artificial doctors"; they are (at best) cognitive augmentation tools. Their potential to improve quality, efficiency, and equity in health is real, but it will only be realized if they are treated like any other medical technology: with specifications for use, independent validation, continuous monitoring, and ethical governance.

Clinical adoption of LLMs must be gradual, supervised, and transparent. Each step must be accompanied by documentation, training, and impact assessment. Trust should not come from the model's reputation, but from locally produced evidence.

The challenge in 2025 is to transform fascinating models into reliable clinical tools: patient-centric, auditable, and fair. This requires collaboration between physicians, engineers, regulators, and patients.

We believe that the future of AI in medicine will not depend on the size of the models, but on the maturity of the institutions that use them.

REFERENCES

1. Thirunavukarasu AJ, Ting DSJ, Elangovan A, et al. Large language models in medicine. Nat Med. 2023;29:1930–1940.
2. Riedemann L, et al. The path forward for large language models in medicine is open. NPJ Digit Med. 2024 Nov 27;7(1):339. doi:10.1038/s41746-024-01344-w.
3. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med. 2024;30:1134–1142. doi:10.1038/s41591-024-02855-5.
4. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med. 2023;183(6):589–596. doi:10.1001/jamainternmed.2023.1838.
5. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv [cs.CL]. 2023 Apr 12. doi:10.48550/arXiv.2303.13375.
6. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. 2025;31:943–950. doi:10.1038/s41591-024-03423-7.
7. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health. 2023 Mar;5(3):e107–e108. doi:10.1016/S2589-7500(23)00021-3. Epub 2023 Feb 6. PMID: 36754724.

8. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. Lancet Digit Health. 2023 Jun;5(6):e333–e335. doi:10.1016/S2589-7500(23)00083-3. Epub 2023 Apr 27. PMID: 37120418.
9. Koller D. Why we support and encourage the use of large language models in NEJM AI submissions. NEJM AI. 2024;1(1). doi:10.1056/AIe2300128.
10. Meng X, Yan X, Zhang K, et al. The application of large language models in medicine: a scoping review. iScience. 2024 Apr 23;27(5):109713. doi:10.1016/j.isci.2024.109713. PMID: 38746668; PMCID: PMC11091685.
11. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large (multimodal) models. Geneva: WHO; 2024.
12. U.S. Food and Drug Administration. Artificial Intelligence/Machine Learning-Based Software as a Medical Device Action Plan. Silver Spring, MD: FDA; 2021.
13. European Union. Artificial Intelligence Act. Regulation (EU) 2024/1689 of the European Parliament and of the Council. Brussels: EU; 2024.
14. Maliha G, Gerke S, Cohen IG, et al. Artificial intelligence and liability in medicine: balancing safety and innovation. Milbank Q. 2021 Sep;99(3):629–647. doi:10.1111/1468-0009.12504. Epub 2021 Apr 6. PMID: 33822422; PMCID: PMC8452365.
15. McCoy LG, Swamy R, Sagar N, et al. Assessment of large language models in clinical reasoning: a novel benchmarking study. NEJM AI. 2025;2(10). doi:10.1056/AIdbp2500120.
16. Han T, Adams LC, Bressem KK, et al. Comparative analysis of multimodal large language model performance on clinical vignette questions. JAMA. 2024 Apr 16;331(15):1320–1321. doi:10.1001/jama.2023.27861. PMID: 38497956; PMCID: PMC10949144.
17. Widner K, Virmani S, Krause J, et al. Lessons learned from translating AI from development to deployment in healthcare. Nat Med. 2023;29:1304–1306. doi:10.1038/s41591-023-02293-9.
18. Omiye JA, Gui H, Rezaei SJ, Zou J, Daneshjou R. Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. Ann Intern Med. 2024 Feb;177(2):210-220. doi:10.7326/M23-2772. Epub 2024 Jan 30

BRIEF GLOSSARY [18]

DESIGNATION	DEFINITION
Bias (in AI)	*Systematic errors in the output of a model due to flawed assumptions in the machine-learning process. This is usually from the data the model are trained on and can also be accentuated in the fine-tuning process.*
Fine-tuning	*Further training a pretrained model on a specific task and adjusting the preexisting parameters to achieve better performance for a particular task.*
Foundation model	*A large-scale neural network model trained on vast data to develop broad learning capabilities, which can be fine-tuned for specific tasks. A foundation model can be fine-tuned to generate reports or answer medical questions.*
Generative artificial intelligence	*Models trained on large data sets that can produce seemingly novel realistic content. This can be audio, visual, or text.*
Large language models (LLMs)	*AI models trained on an enormous amount of text data, that are capable of generating humanlike text and learning relationships between words.*
Multimodal LLMs	*Models capable of processing and generating different types of data, such as text, images, and audio. They are an emerging form of LLM with a wide range of potential applications in medicine.*
Neural networks	*Systems inspired by the neuronal connections in the brain that are capable of learning, recognizing patterns, and making predictions on tasks without explicit programming. They are the building blocks of many modern machine-learning (deep-learning) algorithms.*
Pretraining	*The initial phase of training a model on a large data set before fine-tuning it on a task-specific data set. The parameters are updated in the training process.*

Adapted from Ann Intern Med. doi:10.7326/M23-2772