Embracing AI in Medicine: The Role of Large Language Models in Healthcare

Ethan Szpara, DO

Consider the case of a pediatric patient who presents with recurrent sore throat, erythema, no tonsillar exudates, and a negative viral panel. The patient has a positive group A strep (GAS) PCR swab, however, was recently diagnosed with streptococcal pharyngitis a couple weeks ago and just finished a course of amoxicillin 7 days ago. Is this a false positive PCR test result? What is the next appropriate course of action? What is the overall incidence of failed outpatient therapy of GAS versus recurrence? Is repeat treatment indicated, and what would be the most appropriate second-line therapy based on current evidence and guidelines?

Primary care, emergency medicine, and urgent care clinicians might easily encounter a case like this on their next shift. Even though the elements of this case are routine, it is likely many clinicians would have some uncertainty about what the current evidence and relevant guidelines might suggest is the best course of action. Most clinicians, in such instances, will reach for search tools such as PubMed or UpToDate, however, finding an answer to such a specific question can prove challenging and time-consuming. Knowing this, some may choose instead to curbside a colleague or supervising doctor. But is this best practice?

Artificial intelligence (AI) search tools are permeating clinical medicine. A subset of clinicians has already adopted AI tools as their preferred method of referencing answers to these unique clinical questions. However, is this practice ready for prime time? With so many AI software platforms bursting onto the AI scene, how can clinicians know which ones are trustworthy?

AI has rapidly integrated into many aspects of clinical medicine over recent years.[1] Large language models (LLMs) currently serve as key "engines" powering various AI tools that help answer clinical questions. Widely used examples of LLMs include ChatGPT, Claude, Gemini, and Perplexity. While versatile, these general-use tools may lack reliability in specialized medical contexts. As AI and LLMs are increasingly used by clinicians for diagnostic support, guaranteeing reliability and explainability of AI outputs are critical to ensure AI will deliver on the promise of improving efficiency without adversely affecting patient care or safety.[2,3]

Ethan Szpara, DO, cares for patients at Illinois Emergency Medicine Specialists, serving several Emergency Departments in the Chicagoland area, and is a clinical instructor for the University of Illinois at Chicago Emergency Medicine Program at Little Company of Mary Medical Center.

A Brief History of AI in Healthcare and LLMs

Various forms of AI have been implemented in healthcare conceptually for over half a century with the idea of "machine learning" first emerging in the 1950s. In the 1970s, an AI program called MYCIN, created at Stanford University, was first used to help physicians by suggesting when and which antibiotics to use for certain infections.[4] Early AI was largely rules-based, and due to limited computing power and insufficient available data, the extent to which machine learning could occur was limited. Beginning in the 2000s, the leap in computational power and digitized medical data began allowing for AI systems to develop into much more versatile and powerful tools, which began showing promise in areas such as medical imaging, predictive analytics, and diagnostics.[5] More recently, LLMs have become widely available and affordable. This evolution has prompted accelerated adoption among clinicians seeking tools to alleviate the cognitive and administrative burdens of clinical practice.

Introduction to LLMs: How Do They Work?

LLMs can be thought of as "digital brains," which have developed their unique form of understanding through the process of data training. The models are trained on vast datasets so that they might identify patterns and make probabilistic predictions. When queried, LLMs

generate a response based on the patterns they have been trained to identify. In other words, they are not truly "thinking" in a fully human sense. However, like human intelligence, LLMs use prior experience (ie, data training) to detect patterns and, when identified, predict the most likely outcome.

"Designing LLMs with this specific functionality allows for not only fast, reliable answers to clinical conundrums, but also the opportunity for clinicians to learn about the existence of newer evidence and guidelines."

LLMs engage in natural language processing (NLP) specifically to achieve this. NLP can be subdivided into 2 components: natural language understanding (NLU) and natural language generation (NLG). NLU refers to a model's ability to interpret written human language. It allows AI to extract meaning from text, and based on the context, it can then understand questions being asked by a human user. Alternatively, NLG is the process by which models generate written text that conforms to accepted grammatical and syntax rules and is expected to be meaningful to a human reader. LLMs utilize both NLU and NLG elements, and these processes often occur simultaneously.

LLMs develop their expertise by virtue of data training, which involves exposure to vast datasets of text such as books, scientific articles, and websites. The training data used by LLM developers is a critical decision in shaping the models' potential outputs. Like other types of AI, LLMs rely on the concept of neural networks. Neural networks involve layers of mathematical functions called nodes, which process data through their integrated function. The network can predict patterns with higher accuracy by training on larger amounts of diverse data. As you descend through the layers of the network, each layer and node completes various complex tasks, such as recognizing a piece of a pattern. Each layer of the network builds on that pattern as the network progresses through the layers and learns the data. Ultimately, at the last layer of the network, an output emerges based in probabilistic predictions from the learned patterns that appear to "understand" the information being processed. While LLMs do not understand in a human sense, they can approximate understanding by modeling human language statistically.

Application of LLMs in Modern Medical Practice

While younger individuals historically have been the early adopters of new technologies,[6] clinicians of all specialties and ages are showing interest in the potential applications for AI in various domains of patient care.[7] In 2022, the release of GPT-3.5 and then ChatGPT provided a first glimpse into the vast potential applications of LLMs in clinical practice. This first wave of broadly applicable LLMs, for all their promise, suffered from excessive tendency to exhibit bias and "hallucinate" by fabricating responses in attempt to answer user queries without substantive evidential support.[8] Subsequently, additional AI platforms, such as Claude AI, sought to provide more reliable and bias-free output. Perplexity, another AI platform, was designed to improve explainability by offering citations to statements made in its output. However, this explainability does not necessarily confer reliability as references may include non-peer reviewed publications, such as personal blogs or promotional websites. While the potential for specialized LLMs, particularly in medicine, has been apparent, sophisticated users readily recognize the significant dangers that exist with unreliable and/or biased output.

To address this, our team at OpenEvidence, has developed an AI platform designed for practicing clinicians. The model has been specifically engineered to imitate the more nuanced decision-making process a clinician would follow—much like when choosing primary data or literature to support a clinical decision.

OpenEvidence allows providers to ask clinical questions and receive responses sourced from peer-reviewed literature as well as clinical guidelines, and it provides supporting citations so that clinicians can verify their veracity. This functionality can improve efficiency and promote cognitive off-loading.

As an example, let's take the GAS case introduced above. In less than a minute, the model can address all of the questions raised regarding our pediatric patient with sore throat. Specifically, the model generates the response that treatment failure rates can be seen in approximately 10-20% of GAS cases treated with amoxicillin.[9] It also states that recurrent positive results can be due to residual DNA material, chronic carriage, or true symptomatic recurrence.[10] Combing the individual articles with traditional reading or search functions would take much longer to find the same information. Designing LLMs with this specific functionality allows for

not only fast, reliable answers to clinical conundrums, but also the opportunity for clinicians to learn about the existence of newer evidence and guidelines.

Challenges and Ethical Considerations

Integrating AI, especially LLMs, into clinical practice cannot be pursued without careful consideration for the ethical implications of its use. Understanding that hallucinations can occur and how they can be identified and mitigated is central to the safe clinical application of LLMs. Hallucinations are fabricated outputs and unsubstantiated answers to user questions. They occur for various reasons, but without intentionality in training and LLM design, they can prove difficult to detect and prevent.[8] It is imperative that clinicians are aware of the theoretical risk of hallucinations when choosing to use an LLM to assist in clinical decision making.

Bias is another challenge that AI systems face. If a system is trained on the entire internet, it is important to note that this could include sources that are not factually accurate. Bias exists in all forms of text, and LLMs can incorporate this bias through the process of machine learning, replicating it in their output. In healthcare systems, similar bias can occur if AI is trained on data sets that reflect existing biases of clinicians. For example, if an LLM is trained on data in which an already marginalized group has pain inadequately treated, the model may recommend a suboptimal pain management approach to similar patients, thereby perpetuating existing biases.[11]

There are several ways in which LLM developers may mitigate the risks of bias and hallucinations infiltrating outputs. These areas are subjects of intensive ongoing study and have led to an increased understanding of the importance of data quality and human feedback.[8,12] Fine tuning of models based on high quality expert curated data sets is an important component for the delivery of accurate output from LLMs.

The Future of LLMs in Medicine

The future of LLMs in medicine is promising, but their promise being realized will require vigilance from both those who design the tools and the end users (ie, clinicians). Given the rapid developments in AI, it is imperative that users remain aware of limitations when incorporating LLM output into clinical practice. "Trust but verify," is a mantra that has been used to guide the supervision of generations of medical trainees and is an apt mantra for LLM use by clinicians. I have devoted my time and effort to help the developers and engineers at OpenEvidence improve the platform because I believe in the value this product can provide for a busy clinician seeking to practice evidence-based care. However, with growing adoption and acceptance of LLM use, it is critical to remember that it is incumbent on all of us, the human clinicians making medical decisions for our patients, to exercise good judgment and critical thinking before implementing LLM outputs. LLMs are powerful tools, but it is important to remember that we, the clinicians, are the ones from whom the patient receives care. ■

"Given the rapid developments in AI, it is imperative that users remain aware of limitations when incorporating LLM output into clinical practice."

References

1. AIPRM. AI in Healthcare Statistics. AIPRM; 2024. Accessed April 2, 2025. https://www.aiprm.com/ai-in-healthcare-statistics/
2. Elhaddad M, Hamam S. AI-Driven Clinical Decision Support Systems: An Ongoing Pursuit of Potential. *Cureus*. 2024;16(4):e57728. doi:10.7759/cureus.57728
3. Scispot. AI Diagnostics: Revolutionizing Medical Diagnosis in 2025. Scispot. Published March 20, 2024. Accessed April 1, 2025. https://www.scispot.com/blog/ai-diagnostics-revolutionizing-medical-diagnosis-in-2025
4. Press G. 12 AI Milestones: 4. MYCIN—An Expert System For Infectious Disease Therapy. *Forbes*. April 27, 2020. Accessed April 2, 2025. https://www.forbes.com/sites/gilpress/2020/04/27/12-ai-milestones-4-mycin-an-expert-system-for-infectious-disease-therapy/
5. Keragon Team. When Was AI First Used in Healthcare? The History of AI in Healthcare. *Keragon*. Published February 29, 2024. Accessed April 2, 2025. https://www.keragon.com/blog/history-of-ai-in-healthcare
6. Randstad. Generational Divide: AI Adoption. Randstad USA. Accessed April 2, 2025. https://www.randstadusa.com/business/business-insights/workplace-trends/generational-divide-ai-adoption/
7. American Medical Association. 2 in 3 physicians are using health AI: 78% jump in one year. AMA. Published April 1, 2024. Accessed April 2, 2025. https://www.ama-assn.org/practice-management/digital/2-3-physicians-are-using-health-ai-78-2023
8. Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy. *Nature*. 2024;630:625-630. doi:10.1038/s41586-024-07421-0
9. Gerber MA, Baltimore RS, Eaton CB, Gewitz M, Rowley AH, Shulman ST, Taubert KA. Prevention of rheumatic fever and diagnosis and treatment of acute streptococcal pharyngitis: A scientific statement from the American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee of the Council on Cardiovascular Disease in the Young, the Interdisciplinary Council on Functional Genomics and Translational Biology, and the Interdisciplinary Council on Quality of Care and Outcomes Research: Endorsed by the American Academy of Pediatrics. *Circulation*. 2009;119(11):1541-1551. doi:10.1161/CIRCULATIONAHA.109.191959
10. Randel A, Infectious Disease Society of America. IDSA updates guideline for managing group A streptococcal pharyngitis. *Am Fam Physician*. 2013;88(5):338-340.
11. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
12. Mittermaier M, Raza MM, Kvedar JC. Bias in AI-based models for medical applications: challenges and mitigation strategies. *NPJ Digit Med*. 2023;6:113. doi:10.1038/s41746-023-00858-z