CHEST Critical Care

Implementing Large Language Models in Critical Care

What Is Behind the Delay?

Khalil El Gharib, MD

Large language models are artificial intelligence tools that have emerged in a supportive role in various fields, and we are on the cusp of witnessing their true potential in critical care medicine. In this article, we highlight the anticipated applications of these models in this setting, as well as the barriers that separate us from actualizing their deployment, ranging from pragmatic concerns, such as doubts in clinical decision-making, to ethical trustworthiness, along with means to attempt to mitigate these challenges. CHEST Critical Care 2025; 3(3):100180

KEY WORDS: artificial intelligence; benchmarks; critical care; ethics; large language models

Critical care medicine remains an underprioritized field, and 5 years after the pandemic, dedicated resources are still scarce, and the inequality of global access is surprisingly underrecognized.[1] This compounds workload and leads to occupational burnout, which raises the necessity of searching for alternatives to mitigate the challenges encountered in a milieu where patient acuity is high.

The hallmark of the 21st century has undeniably been the rise of artificial intelligence (AI), and its progress allowed it to rapidly transform many sectors. Large language models (LLMs) are certainly the emblem of this evolution.[2]

LLMs are systems trained on vast amounts of data derived from articles, books, and other internet-based contnet..[3] These models are based on transformer architecture to generate human-like text, answer questions, and perform tasks by using pretrained, finetuned algorithms.[4] They function through conversational interfaces where one enters a question or a task to perform in the model's chatbot and it generates a response; some software and online platforms are already integrating LLMs directly into their features.[5] Many LLMs have been developed over the past couple of years, including Gemini, Claude, and the most famous, ChatGPT, and their integration into the medical field has recently intrigued investigators and trialists.

They are capable of passing medical licensing examinations,[6] helping with clinical documentation and medical coding, and facilitating patient-physician communication,[7] conversing in an empathetic manner comparable to that of human physicians.[8] LLMs can learn from highly curated data sets and can help

ABBREVIATIONS: AI = artificial intelligence; CoT = chain-of-thought; LLM = large language model

AFFILIATIONS: From the Division of Pulmonary & Critical Care Medicine, Rutgers Robert Wood Johnson Medical School, New Brunswick, NJ.

outlook.com

CORRESPONDENCE TO: Khalil El Gharib, MD; email: khalil.gharib@

 DOI: https://doi.org/10.1016/j.chstcc.2025.100180

physicians tailor clinical decision paradigms.[9] They exhibit strong clinical performance even in complex cases,[10,11] raising the question of whether LLMs are to be seen truly as standalones in medical practice[12] and as a sound medical information platform.[13] In critical care, AI models outside LLMs can predict patient deterioration and optimize resource use, which can lead to timely interventions, better patient outcomes, and reduced unit readmission.[14] These predictive analytics are of utmost importance in critical care settings, and LLMs' properties were shown to be superior to common risk stratification approaches,[15] which seemingly stems from their underlying probabilistic architecture allowing them to share some of the human physicians' reasoning capacities.[16] In essence, LLMs exhibit promising potential in critical care; let me illustrate it in some examples: after entering into a model's chatbot a related clinical scenario, I can ask it to:

Analyze the data, then stratify the patient and risk of deterioration: Gemini, how do you think this patient is going to do over the next 4 days in the ICU?

Give evidence-based recommendations: This patient has ventilator-associated pneumonia; according to the most recent guidelines, what antibiotics should I give? And is there anything else I should do?

Elaborate differential diagnoses: Patient's shock state and hypotension have not been responding well to high rates of norepinephrine drip, anything else I should be thinking of?

Automate documentation: The patient has been in the ICU for 10 days now and here is his chart; please summarize his ICU course for proper nursing staff handoff before transfer to medical general wards.

Improve communication: ChatGPT, can you please explain the concept of multidrug-resistant pneumonia to the patient and his family in plain language?

Despite these various benefits, physicians and policymakers have been reluctant to integrate LLMs into critical care practice, and expanding AI research in this domain has been comparably slow. Developed below are the main reasons behind this hesitancy, in addition to suggestions that can help overcome these barriers.

Clinical Decision-Making

LLMs were not initially built to serve medicine and health care, and their direct application intoreal-world clinical scenarios has been questionable. One of the limitations remains the uncertainty of whether LLMs are reliable in clinical decision-making, which is a complex multistep process that involves a model's awareness of the most recent guidelines, information-gathering abilities, and robustness in the prompts that are entered in the chatbot and that also should be flexible to the dynamism of the diagnostic data being synthesized along the way.[17] It was only recently demonstrated that generative pre-trained transformer-4 can improve physician management reasoning compared with conventional resources only,[18] but that was based on the sole randomized clinical trial investigating LLMs in this setting. Its generalizability is uncertain at this point, particularly in critical care, where disease phenotype is much more complicated and where human physicians remain the proprietaries of clinical decision-making. The challenges proposed here can be mitigated by providing refined training data, involving ensemble learning, and integrating LLMs in more randomized clinical trials under standardized conditions. Measured outcomes are under investigation at this point. They can include patient outcomes, but more quantifiable ones can be studied beforehand, such as the ability of LLMs to list differential diagnoses in a critical patient, including the correct disease in an accurate probabilistic method, to suggest appropriate cost-effective diagnostic approaches, and so forth. Once proven, this could increase the clinicians' reliability on these models for diagnostic purposes, potentially transferring their output into daily clinical work.

Training

We are training LLMs with more medical data and injecting an exponentially expanding number of medical text corpora into these models. However, data driven solely by critical care scenarios remains limited, and research in this domain has been narrowed to training machine learning models, outside LLMs, with data derived from the Medical Information Mart for Intensive Care and the electronic ICU data sets. The clinical contexts encountered in critical care settings are different from those we see in general medical wards because more parameters are required to interpret the scenario, that is, hourly or even closer vital signs monitoring, telemetry recordings, relevant nursing and medical staff documentation, and exhaustive yet pertinent laboratory and imaging investigations. Even if the data given to the LLM are the most complete, applying its recommendations, if relevant, should be prompt, because when taking care of the critically ill, every minute counts, and delay in

care can be detrimental prognostically. For all these reasons, obtaining robust critical care data aggregated from multiple centers and creating research networks in this regard is essential to build fine-tuned models clinically focused on the unique critical care population before their safe deployment in the ICU. Training should be provided to the ICU staff as well, and they are expected to develop AI literacy and be provided with the tools needed to recognize LLMs' strengths, limitations, and probabilistic operability[19] through tailored training modules, as well as hands-on workshops and simulations specific to this high-stakes environment.

Prompting

Output from LLMs is consistent regardless of how prompts are structured, as these models are designed to extract equal meaning from all parts of the input without prioritizing any section [20]

Prompt engineering is an evolving discipline that crafts LLM inputs in a way that will allow us to obtain the coveted responses,[21] and various techniques are undergoing testing in different sects; chain-of-thought (CoT) and tree-of-thought are being investigated in medical settings.[22] How can we benefit from them when instructing an LLM to help us with the diagnostic/management paradigms in critical care?

As an example, I can type in the chatbot the following prompt:

"Hey ChatGPT, I have a 68-year-old male patient intubated for acute hypoxic respiratory failure due to ARDS with the following ventilator parameters:

Mode: Volume control
Tidal volume: 400 mL
Respiratory rate: 14 breaths/min
F_{IO_2}: 80%
Positive end-expiratory pressure ¼ 10 cm H_2O

The peak pressure went suddenly from 27 to 50 cm H_2O. How should I proceed?"

That is what we call zero-shot prompting, in which I am asking GPT to perform a task without any examples. In CoT,[23,24] I can reason with the model step-by-step, whereas instead I can ask it in the case above: What do you think is happening? or What might be the reasons for this increase in peak pressure? before proceeding into asking management questions. In tree-of-thought,[25] I guide the model into different reasoning possibilities. In the case again, I would type this question instead: An iatrogenic pneumothorax or a kink in the endotracheal tube might cause it. I can distinguish those by first applying an inspiratory pause and measuring the plateau pressure; I can diagnose a pneumothorax with a chest X-ray as well. ChatGPT, what do you think?

Most recently, a new prompting strategy has shown to be promising to mitigate the problems seen here. It is called self-questioning prompting, in which I express doubts about the model's answers and keep questioning them with the model itself: How sure are you about that? This strategy is more performant than simple or CoT prompting.[26] Employing any of these techniques is tempting in a clinical setting, but it still demands rigorous research on which is most precise and in what clinical scenario. Table 1 defines the different prompting tools to be used in medicine and critical care, along with an example of the reasoning process of each when faced with a clinical scenario of a new pulmonary lobar infiltrate in an intubated patient.

Cost

The global growth in LLM use is accompanied by a parallel increase in development cost, because computing power and diffusion into different sects have a not-so-negligible expenditure.[2] Health care as aforementioned is 1 of the top fields witnessing the potential of LLMs in the workplace, and their integration into critical care practice amongst physicians and medical personnel will compound this cost, particularly when these models become US Food and Drug Administration-approved for use in clinical practice. Hospital administrators also may lean toward setting physicians' expectations into providing care for a higher number of patients, because they are now cooperating with AI in co-managing these patients, which opposes the concept initially suggested of implementing LLMs to lessen clinical workload and linked burden.

Ethical Concerns

The rise of LLMs in medicine was accompanied by ethical concerns, and the crescendo of the development of these models further magnified the necessity of addressing them. Although the topic is quite broad, I will emphasize their potential impact on critical care. First and foremost is the problem of data breaches and unintended disclosure of patient-sensitive information,[27] particularly in the absence of regulatory third parties that impose filtering of

the input being injected into the LLMs from patient identifiers and prevent related information from being communicated to others, especially if access to data sets has become public. LLMs

TABLE 1] Prompting Techniques, Definitions, Examples, and Strengths and Weaknesses Associated With Each

Prompting	Definition	Example	Strengths	Weaknesses
Zero-shot	Provides the model with direct instructions or questions, without any in-context examples of how to perform the task	The patient has been intubated for a week now and has a new left lower lobe infiltrate. Gemini, what should I treat?	(1) Simple and easy to implement (2) Requires no task-specific examples	Less effective for novel or nuanced tasks
Chain-of-thought	Prompts the model to show its reasoning process step-bystep before arriving at the final answer	Gemini, create an algorithm illustrating the differential diagnoses associated with this infiltrate and reason with me in the management options of each.	(1) Improves performance on complex reasoning tasks (2) Reduces error by encouraging structured thinking	(1) Less necessary for simple tasks (2) Requires crafting good reasoning examples (3) Can increase the length of the output
Tree-of-thought	Explores multiple potential reasoning paths in parallel, evaluates them, and backtracks when necessary to find the best solution	Gemini, if I tell you that the patient has fevers and purulent secretions when suctioning inside the endotracheal tube, would this infiltrate more likely represent VAP?	(1) More robust reasoning by exploring multiple options (2) Highly effective for complex tasks (3) Enables error correction through evaluation and backtracking	(1) Most complex to implement (2) Computationally very expensive
Self- questioning	Encourages the model to explicitly ask follow-up questions to break down a complex problem or query into smaller, more manageable subproblems	Are there other possibilities besides the VAP diagnosis that you just suggested?	To be determined	To be determined

VAP ¼ ventilator-associated pneumonia.

also tend to perpetuate societal biases,[28,29] disfavoring underserved and marginalized populations. This raises the question of whether LLMs behave the same with the data obtained in critical care and propagate harmful biases in a setting where diagnostic and management options should be tailored to the patient as a unique entity, with special consideration of all the aspects of the patient's background. The problem can be addressed by establishing reinforcement learning, using feedback from human physicians aware of the distinct demographics that patients come from. The third ethical pitfall is the phenomenon of "hallucinations," in which models fabricate facts and recommendations based on fictitious credentials, which is quite dangerous if related output is directly applied to managing patients in the ICU without human oversight of the credibility of resources used. Approximately 5 years into the public availability of LLMs, the problem with hallucinations remains unresolved, because they stem from a complex genesis that human control is by itself insufficient.[30]

Thus, making sure that LLMs overcome the aforementioned ethical obstacles is not an easy task, but it is a must before deploying this technology in the daily critical care workflow and to be certain that they are developed under the 4 pillars of medical ethics (beneficence, autonomy, nonmaleficence, and justice).

Conclusion

Minimizing cognitive load is becoming increasingly appreciated among physicians in the ICU,[31] and LLMs seem promising not only in daily clinical practice but also in medical education and research.[32,33] However, we should overcome several challenges before safe implementation of these AI tools in critical care is possible.[14] Research here should be more laborious to strengthen model reliability and interpretability, reduce errors in prompt architecture, and favor model robustness.[34] Critically ill patients are most likely to benefit from the transformative potential of LLMs in medicine, but we cannot witness that before ensuring AI is transparent in its applications herein and models' performance surpasses, or at least meets, established benchmarks.

Financial/Nonfinancial Disclosures

None declared

Acknowledgments

Author contributions: K. E. G.: writing—original draft and editing

References

1. Crawford AM, Shiferaw AA, Ntambwe P, et al. Global critical care: a call to action. Crit Care. 2023 Jan 20;27(1):28.
2. The Economist. Large language models are getting bigger and better.The Economist website. Accessed February 20, 2025. https://www.economist.com/science-and-technology/2024/04/17/large-languagemodels-are-getting-bigger-and-better
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930-1940.
4. Kasneci E, Seßler K, Küchemann S, et al. Chatgpt for good? on opportunities and challenges of large language models for education. Learning and Individual Differences. 2023;103:102274.
5. Revelo. How to Integrate Large Language Models (LLMs), Into YourProduct, Revelo website. Accessed February 20, 2025. https://www.revelo.com/blog/large-language-models
6. Chen Y, Huang X, Yang F, et al. Performance of ChatGPT and Bard on the medical licensing examinations varies across different cultures: a comparison study. BMC Med Educ. 2024;24(1):1372.
7. Generative AI for Clinical Conversations. Abridge website. AccessedMarch 12, 2025. https://www.abridge.com/
8. Tu T, Palepu A, Schaekermann M, et al. Towards conversational diagnostic AI. Nature. 2025;642:442-450.
9. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large language models in medicine. Commun Med (London). 2023;3(1):141.
10. Buckley TA, Crowe B, Abdulnour R-EE, Rodman A, Manrai AK. Comparison of frontier open-source and proprietary large language models for complex diagnoses. JAMA Health Forum. 2025;6(3): e250040.
11. Eriksen AV, Möller S, Ryg J. Use of GPT-4 to diagnose complexclinical cases. N Engl J Med AI. 2023;1(1). https://doi.org/10.1056/ AIp2300031
12. Goh E, Gallo R, Hom J, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 2024;7(10):e2440969.
13. Wu V, Casauay J. OpenEvidence. Fam Med. 2025;57(3):232-233.
14. Spencer E-J, Economou-Zavlanos NJ, van Genderen ME. What if we do, but what if we don't? The opportunity cost of artificial intelligence hesitancy in the intensive care unit. Intensive Care Med. 2025;51(2):378-381.
15. Jentzer JC, Kashou AH, Murphree DH. Clinical applications of artificial intelligence and machine learning in the modern cardiac intensive care unit. Intelligence-Based Medicine. 2023;7: 100089.
16. Restrepo D, Rodman A, Abdulnour R-E. Conversations on reasoning: large language models in diagnosis. J Hosp Med. 2024;19(8):731-735.
17. Hager P, Jungmann F, Holland R, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med. 2024;30(9):2613-2622.
18. Goh E, Gallo RJ, Strong E, et al. GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial. Nat Med. 2025;31(4):1233-1238.
19. Cecconi M, Greco M, Shickel B, Vincent J-L, Bihorac A. Artificial intelligence in acute medicine: a call to action. Crit Care. 2024;28(1): 258.
20. Liu T, Duan Y. Beware the self-fulfilling prophecy: enhancing clinical decision-making with AI. Crit Care. 2024;28(1):276.
21. Marvin G, Hellen N, Nakatumba-Nabende J. Prompt Engineering inLarge Language Models. In: Jacob IJ, Piramuthu S, FalkowskiGilski P, eds. Data Intelligence and Cognitive Informatics. ICDICI 2023. Algorithms for Intelligent Systems. Singapore: Springer; 2024. https://doi.org/10.1007/978-981-99-7962-2_30

22. Yuan M, Bao P, Yuan J, et al. Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant. Medicine Plus. 2024;1(2):100030.

23. Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. NPJ Digit Med. 2024;7(1):20.

24. Zhang Z, Ni H. Critical care studies using large language models based on electronic healthcare records: a technical note. J Intensive Med. 2024;5(2):137-150.

25. Ji Y, Yu Z, Wang Y. Assertion detection in clinical natural language processing using large language models. 2024;2024:242-247.

26. Wang Y, Zhao Y, Petzold L. Are large language models ready for healthcare? a comparative study on clinical language understanding. Proceedings of the 8th Machine Learning for Healthcare Conference. 2023;219:804-823.

27. Ong JCL, Chang SY-H, William W, et al. Medical ethics of largelanguage models in medicine. N Engl J Med AI. 2024;1(7). https://doi.org/10.1056/AIra2400038

28. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. Lancet Digit Health. 2024;6(1): e12-e22.

29. Hagendorff T, Danks D. Ethical and methodological challenges in building morally informed AI systems. AI Ethics. 2023;3(2):553-566.

30. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of chatgpt and other large language models. JAMA. 2023;330(4):315-316.

31. Held N, Neumeier A, Amass T, et al. Extraneous load, patient census, and patient acuity correlate with cognitive load during ICU rounds. Chest. 2024;165(6):1448-1457.

32. Salvagno M, Taccone FS. Artificial intelligence is the new chief editor of Critical Care (maybe?). Crit Care. 2023;27(1):270.

33. Furfaro D, Celi LA, Schwartzstein RM. Artificial intelligence in medical education: a long way to go. Chest. 2024;165(4):771-774.

34. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow. J Med Internet Res. 2023;25: e48659.