Journal of Intensive Medicine 5 (2025) 113–118

Contents lists available at ScienceDirect

Journal of Intensive Medicine

journal homepage: www.elsevier.com/locate/jointm

Review

Large language models in critical care

Laurens A. Biesheuvel[1,#], Jessica D. Workum[2,3,#], Merijn Reuland[1], Michel E. van Genderen[3], Patrick Thoral[1], Dave Dongelmans[4], Paul Elbers[1,*]

[1] Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence, Amsterdam Medical Data Science, Amsterdam Public Health, Amsterdam Institute for Immunity and Infectious Diseases, Amsterdam Cardiovascular Science, Amsterdam UMC, Vrije Universiteit, University of Amsterdam, Amsterdam, The Netherlands
[2] Department of Intensive Care, Elisabeth-TweeSteden Hospital, Tilburg, The Netherlands
[3] Department of Adult Intensive Care, Erasmus Medical Center, Rotterdam, The Netherlands
[4] Department of Intensive Care Medicine, Amsterdam UMC, National Intensive Care Evaluation (NICE) Foundation, Amsterdam UMC location University of Amsterdam, Department of Medical Informatics, Amsterdam Public Health, Amsterdam, The Netherlands

ARTICLE INFO

Editor: Jingling Bao/Zhiyu Wang

Keywords:
Large language models
Intensive care medicine
Critical care medicine
Natural language processing
Artificial intelligence
Machine learning

ABSTRACT

The advent of chat generative pre-trained transformer (ChatGPT) and large language models (LLMs) has revolutionized natural language processing (NLP). These models possess unprecedented capabilities in understanding and generating human-like language. This breakthrough holds significant promise for critical care medicine, where unstructured data and complex clinical information are abundant. Key applications of LLMs in this field include administrative support through automated documentation and patient chart summarization; clinical decision support by assisting in diagnostics and treatment planning; personalized communication to enhance patient and family understanding; and improving data quality by extracting insights from unstructured clinical notes. Despite these opportunities, challenges such as the risk of generating inaccurate or biased information "hallucinations", ethical considerations, and the need for clinician artificial intelligence (AI) literacy must be addressed. Integrating LLMs with traditional machine learning models – an approach known as Hybrid AI – combines the strengths of both technologies while mitigating their limitations. Careful implementation, regulatory compliance, and ongoing validation are essential to ensure that LLMs enhance patient care rather than hinder it. LLMs have the potential to transform critical care practices, but integrating them requires caution. Responsible use and thorough clinician training are crucial to fully realize their benefits.

Introduction

The introduction of chat generative pre-trained transformer (ChatGPT) has sparked an unprecedented interest in the possibilities of natural language processing (NLP), a subfield of artificial intelligence (AI). For the first time, we witness systems that are capable of processing data and generating highly accurate human-like outputs. Unlike traditional machine learning (ML) models, which are typically designed for single, narrow tasks, large language models (LLMs) excel in their ability to perform multiple diverse functions within a single framework. This versatility enables applications ranging from natural language understanding to multimodal data processing. As these systems are becoming increasingly powerful, this breakthrough promises significant implications for any discipline, medical or non-medical. For critical care, these developments could open new possibilities for enhancing patient care and clinical workflows. The rapid pace of advancement in this field raises the question if these developments will transform the landscape of critical care. Therefore, this article aims to set out the history, current applications, limitations, and future prospects of LLMs for critical care medicine.

History of LLMs

As the name suggests, LLMs are characterized by their sizes. Both in terms of the massive datasets they are trained on, as well as the extensive number of numerical parameters (often billions) that are involved to process and generate human language in a multitude of languages. As language contains information, after

* Corresponding author: Paul Elbers, Department of Intensive Care Medicine, Amsterdam UMC, De Boelelaan 1117, 1081 HV, Amsterdam, The Netherlands.
E-mail address: p.elbers@amsterdamumc.nl (P. Elbers).
# Laurens Biesheuvel and Jessica Workum contributed equally to this work.

https://doi.org/10.1016/j.jointm.2024.12.001
Received 7 November 2024; Received in revised form 29 November 2024; Accepted 1 December 2024
Available online 24 December 2024

training, these parameters seem to encode extensive knowledge. Most LLMs rely on a deep learning network architecture known as *transformers*, which were introduced in a revolutionary paper by Vaswani et al.[1] in 2019. This marked the start of a rapidly evolving set of highly proficient and scalable NLP models.[2] Most notably, this resulted in the introduction of ChatGPT in November 2022,[3] a novel way of interacting with the underlying transformer LLM for public users, who had not yet been familiar with this technology. Since then, the technology has further evolved and is seeing more and more useful applications, beyond mere chatting with an AI chatbot, such as AI-powered automation. For example, LLMs are used for large-scale document processing, automated data extraction, summarization, content creation, and enhancing search engine capabilities. In addition to processing text, transformers can be set up to process multi-modal input, such as images, sensory data, or lab results.[4,5] Even more, some of the most recent LLMs are capable of outputting audio based on real-time audio input, rather than text in-text out. As a result, it is possible to communicate with AI as if the user is communicating through the phone.[6] While still in its early stages, this opens new doors for more interactive ways of utilizing LLMs.

LLMs in Healthcare

As LLMs excel at processing unstructured data, they are potentially well-suited to healthcare settings. Clinical notes, patient histories, and medical literature are abundant in healthcare and, until recently, were not effectively accessible through traditional data science methods. Because LLMs are trained in a diverse set of various languages, they contain a massive amount of inherent general knowledge. However, as medical jargon is a subset of various languages, it could be argued that general-purpose LLMs might not be completely capable of capturing certain intricacies in medicine. The understanding of nuanced clinical language, treatment guidelines, and rare medical conditions often requires specialized contextual knowledge that early general-purpose LLMs seemed to lack. To mitigate this limitation of general-purpose LLMs, specialist LLMs were tailored for healthcare by specifically training them on medical texts, including clinical guidelines, medical literature, and patient notes. A notable example includes Med-PaLM Multimodal (Med-PaLM M),[7] which was introduced in July of 2023 by Google. It is a single model that is capable of performing a multitude of tasks, including medical question answering, image interpretation, and radiology report generation. Impressively, it has shown to reach near state-of-the-art performance on all 14 tasks included in their benchmark. It is however important to note that general-purpose LLMs are advancing at a pace that exceeds the development of specialized healthcare models, with impressive results for healthcare applications. For example, the GPT-4 model passes the United States Medical Licensing Examination (USMLE) with an overall average score of 86.7%.[8] Similar scores have been seen in more specialized fields such as Nephrology[9] and Oncology.[10]

Use Cases

The use cases for critical care, and healthcare in general, are myriad. To illustrate this, a major electronic health record (EHR) vendor is currently working on 60 different AI implementations for their EHR system, including LLM-based implementations, some of which are already commercially available.[11] LLM applications can be categorized into their intended use (e.g., administrative support, clinical decision support [CDS], personalizing communication, logistical aid, and improving data quality) and the intended end user (e.g., clinician, patient, manager, and researcher). From both a legal and implementation standpoint, for each LLM application, it is important to have a clear notion of its intended use and the intended end user. More importantly for each application validation and effects due to various biases and maintenance of the aforementioned must be clear. We will now discuss various LLM applications that could be of potential value in critical care medicine.

Administrative support

LLMs show significant potential in reducing the administrative workload for clinicians by streamlining and optimizing workflow processes. The following examples illustrate how LLMs can aid clinicians in their administrative tasks. First, by processing patient data and clinical notes almost instantly, LLMs can generate concise and accurate summaries of patient stays.[12] A recent study (that has not yet been peer reviewed) shows that LLM-generated summaries were non-inferior to physician-written summaries in terms of completeness and correctness while being 28 times faster than the clinician.[13] Clinical note summarization can be expanded by adding laboratory results, radiology reports, and vitals. Therefore, patient chart summarization can be particularly advantageous for prolonged patient stays with cluttered documentation. This ensures that essential yet scattered information from the patient chart is collected in one summary and communicated effectively to subsequent healthcare providers. Another useful LLM application is LLM-based chart search, where information can be quickly extracted by answering specific queries by healthcare professionals.[14] For instance, a clinician may need a rapid overview of a patient's recent lab results, medication changes, or notable events during their hospital stay. By querying the LLM, the clinician can retrieve this information immediately, which can be extremely valuable in urgent critical care situations. Furthermore, LLMs seem to be particularly useful in reducing the administrative burden for clinicians when combined with audio recordings and a transcription service.[15] This is called *ambient listening*. In the intensive care unit (ICU), a plethora of potential use cases can be considered valuable, for example, during rounds, during family conversations, or during multidisciplinary meetings. The first stage of ambient listening maturation in clinical practice would be automatic structured documentation. The second stage would be to combine this technology with automatic administrative actions or ordering.[16] When ambient listening technology is implemented in clinical practice, it is paramount that it is thoroughly clinically tested and validated, and that the clinician is always obliged to review and confirm the prepared notes and orders before they are made part of the patient chart.

CDS

In addition to optimizing workflows, LLMs show potential in assisting clinicians with medical reasoning, where they sug-

gest differential diagnoses, diagnostic tests, or treatment options based on physician input. LLMs have been shown to perform similarly to physicians in diagnostic accuracy and clinical reasoning,[17] or even outperform clinicians in making rare and complex diagnoses.[18] These capabilities can extend to providing diagnostic insights by integrating patient data and medical knowledge, offering treatment suggestions aligned with evidence-based guidelines, and supporting prognostic predictions. When using AI for CDS, a concern with typical deep learning models is that they are often perceived as *black box* models due to their inherent complex architecture that compromises explainability.[19] However, LLMs can be prompted to provide reasoning to make the output interpretable,[20] which may reduce these concerns by enhancing transparency in the model's responses.

However, it must be noted that when an LLM is used to assist in diagnosing or treatment, it could be considered a medical device, and therefore, could fall under specific legislation such as the Food and Drug Administration (FDA) in the United States or the Medical Device Regulation (MDR) in Europe. This prohibits the tool's use until it has passed a comprehensive evaluation to confirm it meets specific quality standards, including clinical proof of safety and efficacy, which often necessitates randomized controlled trials (RCTs). This requirement limits the rapid adoption of LLMs for CDS in healthcare. Additionally, when using publicly accessible LLMs such as ChatGPT with medical data, it could result in data leakage and therefore, breaching doctor-patient confidentiality. Therefore, caution is advised.

Personalized communication

LLM-based chatbots have the capability to adapt their language to the specific user's needs. Notably, their output has demonstrated a level of empathy that exceeds that of clinicians.[21,22] These tools could therefore be utilized to improve patient communication and understanding, for example, by translating medical jargon into layman's terms and adapting the language as needed, based on the input of the user. An example of this is the generation of a supplemental discharge summary in patient-friendly language.[23] Additionally, their interactive nature allows patients to ask questions, further enhancing their understanding of complex medical information. In healthcare, such adaptability could significantly enhance health literacy among patients. In critical care, specifically, patients and families often experience a lack of clarity in communication regarding diagnosis and prognosis[24] by their healthcare providers and express a need for more detailed and understandable communication. LLM-based chatbots could offer a solution by providing clear and accessible explanations of complex medical information, in an interactive way.

Logistical aid

LLMs in the ICU can also be useful outside of individual patient care. Outside of healthcare, LLMs have been used for complex supply chain management by using agents.[25] Agents are autonomous software entities capable of performing complex multimodal tasks independently or collaboratively within a system. Although actual use in healthcare is still limited, the potential for LLM-based agents in critical care logistics could be significant. For instance, LLM-based agents could be used to dynamically predict ICU bed availability, coordinate the distribution of supplies such as medication, and adjust staffing levels based on patient acuity and forecasted demand. Traditional ML models are typically designed for narrow tasks, have limited adaptability, and often require retraining. In contrast, LLM-based agents are more flexible and are able to handle diverse and multimodal tasks in real time potentially making them better suited for the dynamic ICU environment. However, no studies have yet been performed on the application of LLM-based agents in critical care.

Improving data quality

Due to the high amount of monitoring and high frequency of measurements that is inherent to the ICU, there is a high amount of structured data continuously being produced in critical care. While this abundance of data has already been shown to enable big data research, there remains a large portion of unused high-quality data in the form of unstructured patient notes. These notes contain intricacies that are not available in structured data, such as detailed patient history, diagnostic reports, and treatment plans. In addition, subtle changes in a patient's condition may not be sufficiently captured in structured data alone. Furthermore, these notes can give important context in interpreting the structured data.[26] Until now, there was no viable way to make use of this data in a way that is manageable and scalable. With the advent of LLMs, this may change, as LLMs have the capacity to process and extract potentially relevant elements from these notes efficiently and at scale. When this data is effectively used, the implications for research and personalized medicine are considerable. By integrating data from unstructured notes, predictive models could more effectively alert potential adverse events that may otherwise go unnoticed.[27] Furthermore, by utilizing LLMs to improve data registration and quality, this would provide feedback on the predictive AI models already utilized in the ICU. Therefore, this could then also improve the accuracy of discriminative ML, such as prediction models and classifiers. Additionally, LLMs can be harnessed for data harmonization across institutions, enabling big data research at scale.[28] Indirectly, these improvements in data research can enhance the capacity to deliver personalized medicine.

Limitations and Considerations

Although LLMs have a high potential to have a transformative impact on the workflow of critical care in the short term, there remain caveats that should be taken into account that are valid throughout healthcare. Despite its potential, it is important to recognize that this technology is new, and as with any new advancement, its implications for clinical practice remain largely speculative. The added benefit of LLMs in current medical practice is yet to be established as clinical validation studies are lacking. The absence of peer-reviewed evidence that demonstrates the efficacy of LLMs in clinical practice, along with current limitations, underscores that implementation in a high-risk environment such as healthcare and critical care in particular should be approached with caution and careful consideration.

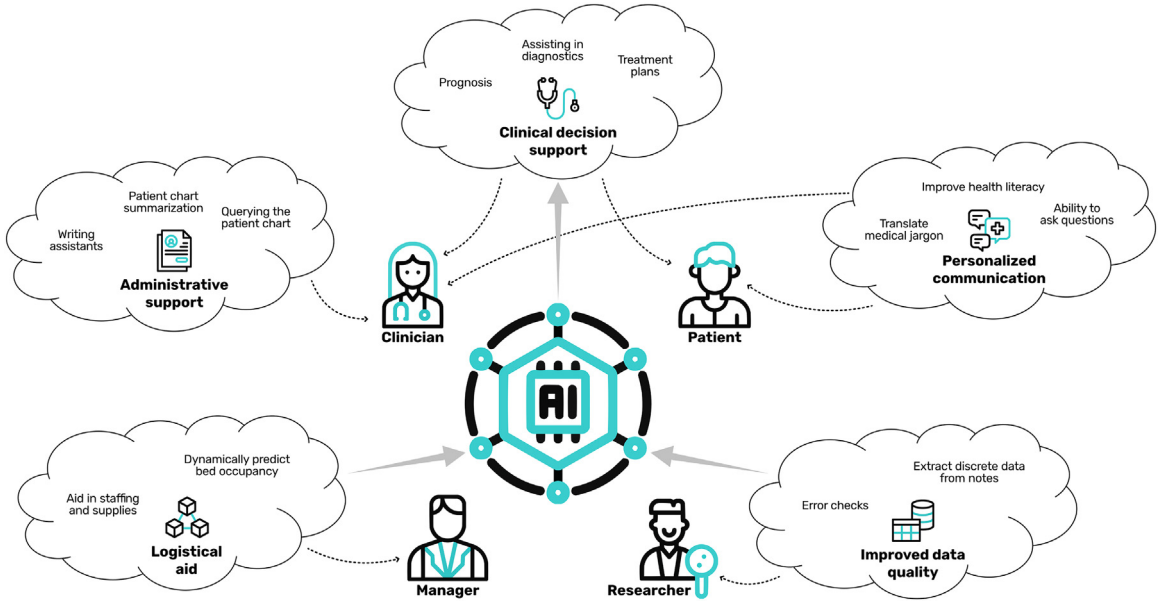

Figure 1. LLMs or hybrid AI support critical care by assisting clinicians, patients, managers, and researchers. Key uses include administrative support, decision-making aid, personalized communication, logistical planning, and data quality enhancement.
AI: Artificial intelligence; LLMs: Large language models.

A notable concern is the tendency of LLMs to generate fabricated responses with confidence, often termed "hallucinations." LLMs can hallucinate because they generate responses based on learned response patterns rather than validated factual knowledge. When generating, they predict the probability of the next word in a sequence, using the probabilities that were derived from the training data. While being exposed to a massive amount of text data from various resources, including books and internet pages, they may still generate plausible-sounding but false information when generating content addressing topics for which the model's training data offers limited representation. While this is particularly concerning for healthcare, where false outputs can have serious implications, hallucinations can be mitigated using techniques such as retrieval augmented generation (RAG), where LLMs are combined with external knowledge bases. By letting LLMs query and analyze these external sources in real-time and formulating responses solely based on the information retrieved from the external knowledge base, they are not only able to generate output that is more accurate but also properly referenced. In healthcare, these sources can be institutional protocols, international guidelines, or scientific medical literature. While such techniques may significantly reduce hallucinations, they remain a valid concern that healthcare providers should be aware of. Extensive testing, clinical validation, and continuous monitoring are necessary to verify reliability before and during deployment.

Another limitation of LLMs that is paramount to consider is that there may exist a tendency of these models to produce biased output.[29] These are the result of imbalances or existing biases in the training data that the model learns from.[30] For models such as GPT-4 that have not fully disclosed model characteristics, the extent of imbalance is difficult to assess due to a lack of transparency about training data.[31] A study by Zack et al.[30] evaluated racial or gender biases in healthcare for the GPT-4 model. They found that the LLM propagated or even amplified societal biases. When GPT-4 was prompted to produce clinical vignettes, they found that the LLM consistently stereotyped demographic presentations for a multitude of diseases. Also, when generating differential diagnoses, it included diagnoses that reflected stereotypes associated with specific ethnicities and genders. Furthermore, their results showed that there was an association between demographic characteristics and recommendations for relatively expensive procedures. These results indicate that using these models for CDS could lead to inequities in care and potentially skew clinical judgment, ultimately posing significant risks to patient safety. For the clinician, caution and careful interpretation of outputs are advised. However, as these produced biases may not be immediately evident to the clinician, adequate oversight and mitigation strategies are crucial. These could include correcting for bias during the training process, fine-tuning models with representative clinical data, incorporating bias detection mechanisms, and conducting ongoing bias audits on model outputs.

Additionally, as a requirement in Article 4 of the EU Artificial Intelligence Act[32] for implementation in practice, it is mandated that those involved in the application and use of AI systems in healthcare – such as clinicians – are sufficiently trained in AI (AI literacy) to ensure safe and effective use. Training must be aligned with the technical expertise, experience, and clinical environments of healthcare professionals. Specific training programs in AI for healthcare providers could help them use LLMs responsibly, understand their limits, and follow best practices for safe use.

There are many other significant caveats to consider, including the potential generation of harmful content,[31] privacy concerns, EHR integration challenges, cost of resources, environmental footprint, and regulatory standards. Due to these challenges, a level of caution is required when utilizing LLMs. Moreover, it is important to note that, contrary to typical ML models, LLMs are not deterministic by design. They generate responses using probabilistic sampling methods rather than selecting only the single most likely outcome. This design choice allows them

to avoid repetitive or overly rigid answers by exploring a range of plausible responses learned from patterns in training data. However, this sampling-driven variation can lead to differing interpretations from similar inputs. This probabilistic nature of LLMs mandates that providers and deployers of these tools need to continuously monitor output and process user feedback to address unexpected or inappropriate responses. Furthermore, with each update of the LLMs, clinical validation per use case should be repeated. This requires tremendous diligence when implementing LLMs into critical care. Healthcare deployers (i.e. hospitals) need to have a long-term vision and governance strategy in place to aid in the safe implementation of LLMs in healthcare.

Hybrid AI

Hybrid AI is an emerging trend in healthcare that combines the strengths of probabilistic generative AI with deterministic ML models while mitigating their respective limitations. In this approach, ML models handle the complex tasks of analyzing symptoms, diagnostic tests, and patient history to facilitate accurate diagnoses. Simultaneously, LLMs could translate these results into clear, comprehensible language and provide additional medical information. This could furthermore be expanded with an interactive component, enabling the user to ask questions. By creating synergy between different AI models, the Hybrid AI approach aims to enhance decision-making accuracy, improve communication, and mitigate issues like hallucinations and bias, ultimately resulting in more reliable and transparent patient care.

Conclusion

The introduction of LLMs in healthcare is showing vast promise for various applications for the clinician, such as administrative support and CDS, for the patients, such as personalized communication and improving health literacy, for logistics and data quality. Their immediate impact is likely to be seen in reducing the administrative burden for the clinician (Figure 1). However, as LLMs are becoming more and more capable, the prospect of LLMs serving as interactive virtual medical experts, analyzing data in real time, and providing insights while considering the latest clinical guidelines and research may not be limited to science fiction in the coming future. While their potential is clear, there remain challenges that cannot be overlooked. Validation studies are essential to ensure the accuracy and reliability of LLMs in various clinical scenarios, and training healthcare staff to effectively use LLMs is essential to mitigate potential harmful responses. Hybrid AI might combine the best of both worlds, using traditional ML for CDS combined with the NLP capabilities and adaptability of generative AI for improving explainability, compiling scattered information from the patient chart, and retrieving the latest medical information from various validated sources. LLMs therefore show vast promise in allowing healthcare providers to focus more on what matters: providing care.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the author(s) used ChatGPT-4o in order to improve readability and language. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT Authorship Contribution Statement

Laurens A. Biesheuvel: Writing – review & editing, Writing – original draft, Conceptualization. **Jessica D. Workum:** Writing – review & editing, Conceptualization. **Merijn Reuland:** Writing – review & editing. **Michel E. van Genderen:** Writing – review & editing. **Patrick Thoral:** Writing – review & editing. **Dave Dongelmans:** Writing – review & editing. **Paul Elbers:** Writing – review & editing, Supervision, Conceptualization.

Acknowledgments

None.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics Statement

Not applicable.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. interest.

Data Availability

The data sets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:1706.03762 2017.
[2] Cascella M, Semeraro F, Montomoli J, Bellini V, Piazza O, Bignami E. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. J Med Syst 2024;48(1):22. doi:10.1007/s10916-024-02045-3.
[3] Introducing ChatGPT. Available from: https://openai.com/index/chatgpt/. [Accessed November 06, 2024].
[4] Meskó B. The impact of multimodal large language models on health care's future. J Med Internet Res 2023;25:e52865. doi:10.2196/52865.
[5] Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: a survey. IEEE Trans Pattern Anal Mach Intell 2023;45(10):12113–32. doi:10.1109/tpami.2023.3275156.
[6] ChatGPT can now see, hear, and speak. Available from: https://openai.com/index/chatgpt-can-now-see-hear-and-speak/. [Accessed November 06, 2024]
[7] Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang P.-C, et al. Towards generalist biomedical AI. arXiv:2307.14334 2023.
[8] Nori H., King N., McKinney S.M., Carignan D., Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375 2023.
[9] Wu S, Koo M, Blum L, Black A, Kao L, Fei Z, et al. Benchmarking open-source large language models, GPT-4 and Claude 2 on multiple-choice questions in nephrology. NEJM AI 2024;1(2). doi:10.1056/AIdbp2300092.
[10] Rydzewski NR, Dinakaran D, Zhao SG, Ruppin E, Turkbey B, Citrin DE, et al. Comparative evaluation of LLMs in clinical oncology. NEJM AI 2024;1(5). doi:10.1056/aioa2300151.
[11] Artificial Intelligence. Epic. Available from: https://www.epic.com/software/ai/ [Last accessed on 2024 December 17].
[12] Madden MG, McNicholas BA, Laffey JG. Assessing the usefulness of a large language model to query and summarize unstructured medical notes in intensive care. Intensive Care Med 2023;49(8):1018–20. doi:10.1007/s00134-023-07128-2.

[13] Schoonbeek R, Workum J, Schuit S, Doornberg J, Laan T, Bootsma-Robroeks C. Completeness, correctness and conciseness of physician-written versus large language model generated patient summaries integrated in electronic health records. 2024. doi: 10.2139/ssrn.4835935.
[14] Ahsan H, McInerney DJ, Kim J, Potter C, Young G, Amir S, et al. Retrieving evidence from EHRs with LLMs: possibilities and challenges. Proc Mach Learn Res 2024;248:489–505.
[15] Barr PJ, Gramling R, Vosoughi S. Preparing for the widespread adoption of clinic visit recording. NEJM AI 2024;1(11):AI2400392. doi:10.1056/AIp2400392.
[16] Seth P, Carretas R, Rudzicz F. The utility and implications of ambient scribes in primary care. JMIR AI 2024;3:e57673. doi:10.2196/57673.
[17] Cabral S, Restrepo D, Kanjee Z, Wilson P, Crowe B, Abdulnour R-E, et al. Clinical reasoning of a generative artificial intelligence model compared with physicians. JAMA Intern Med 2024;184(5):581–3. doi:10.1001/jamainternmed.2024.0295.
[18] Abdullahi T, Singh R, Eickhoff C. Learning to make rare and complex diagnoses with generative ai assistance: qualitative study of popular large language models. JMIR Med Educ 2024;10:e51391. doi:10.2196/51391.
[19] Okada Y, Ning Y, Ong MEH. Explainable artificial intelligence in emergency medicine: an overview. Clin Exp Emerg Med 2023;10(4):354–62. doi:10.15441/ceem.23.145.
[20] 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. doi:10.1038/s41746-024-01010-1.
[21] Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence Chabot responses to patient questions posted to a public social media forum. JAMA Intern Med 2023;183(6):589–96. doi:10.1001/jamainternmed.2023.1838.
[22] Luo M, Warren CJ, Cheng L, Abdul-Muhsin HM, Banerjee I. Assessing empathy in large language models with real-world physician-patient interactions. arXiv:2405.16402 2024.
[23] Zaretsky J, Kim JM, Baskharoun S, Zhao Y, Austrian J, Aphinyanaphongs Y, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Netw Open 2024;7(3) e240357–le24035e. doi:10.1001/jamanetworkopen.2024.0357.
[24] Cezar AGd, Castanhel FD, Grosseman S. Needs of family members of patients in intensive care and their perception of medical communication. Crit Care Sci 2023;35(1):73–83. doi:10.5935/2965-2774.20230374-en.
[25] Xu L, Almahri S, Mak S, Brintrup A. Multi-agent systems and foundation models enable autonomous supply chains: opportunities and challenges. IFAC-PapersOnLine 2024;58(19):795–800. doi:10.1016/j.ifacol.2024.09.200.
[26] Culliton P, Levinson M, Ehresman A, Wherry J, Steingrub JS, Gallant SI. Predicting severe sepsis using text from the electronic health record. arXiv:1711.11536 2017.
[27] Murphy RM, Klopotowska JE, de Keizer NF, Jager KJ, Leopold JH, Dongelmans DA, et al. Adverse drug event detection using natural language processing: a scoping review of supervised learning methods. PLoS One 2023;18(1):e0279842. doi:10.1371/journal.pone.0279842.
[28] Dam TA, Fleuren LM, Roggeveen LF, Otten M, Biesheuvel L, Jagesar AR, et al. Augmented intelligence facilitates concept mapping across different electronic health records. Int J Med Inform 2023;179:105233. doi:10.1016/j.ijmedinf.2023.105233.
[29] Kotek H, Dockum R, Sun D. Gender bias and stereotypes in large language models. In: Proceedings of the ACM Collective Intelligence Conference. Delft, Netherlands: Association for Computing Machinery; 2023. p. 12–24. doi:10.1145/3582269.3615599.
[30] Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, Gichoya J, 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):e12–22. doi:10.1016/S2589-7500(23)00225-X.
[31] Gallifant J, Fiske A, Levites Strekalova YA, Osorio-Valencia JS, Parke R, Mwavu R, et al. Peer review of GPT-4 technical report and systems card. PLOS Digit Health 2024;3(1):e0000417. doi:10.1371/journal.pdig.0000417.
[32] Article 4: AI literacy - EU artificial intelligence act. Available from: https://artificialintelligenceact.eu/article/4/[Accessed November 06, 2024].