BMC Medical Education

REVIEW

Open Access

Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: a narrative review

Liam G. McCoy[1], Faye Yu Ci Ng[2], Christopher M. Sauer[3,4*], Katelyn Edelwina Yap Legaspi[5,6], Bhav Jain[4], Jack Gallifant[4,7], Michael McClurkin[8], Alessandro Hammond[9,10], Deirdre Goode[11], Judy Gichoya[12] and Leo Anthony Celi[4]

Abstract

Reports of Large Language Models (LLMs) passing board examinations have spurred medical enthusiasm for their clinical integration. Through a narrative review, we reflect upon the skill shifts necessary for clinicians to succeed in an LLM-enabled world, achieving benefits while minimizing risks. We suggest how medical education must evolve to prepare clinicians capable of navigating human-AI systems.

Keywords Language Model, Medical Education, Technology, Ethics

Introduction

Recent advances in large language model (LLM) technology have raised excitement about their possible role in clinical practice. Even general-purpose models such as OpenAI's ChatGPT have shown significant promise in these applications, as perhaps best captured in a highly publicized paper wherein the model successfully passed the United States Medical Licensing Examinations (USMLE) [1]. Results have been even more impressive for domain-specific models, such as Google's Med-PaLM2 or GatorTron [2], which answered a wide range of medical questions at an expert level [3]. These striking results raise an important question: are medical education systems adequately preparing the next generation of clinicians to work alongside these models?

Within this article, we offer a brief outline of recent LLM progress as it relates to healthcare, and seek to envision how these models—alongside other artificial intelligence (AI) technologies-–may shift the nature of clinical practice. We critically examine the clinical competencies

*Correspondence:
Christopher M. Sauer
sauerc@mit.edu
[1]Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada
[2]Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[3]Institute for Artificial Intelligence in Medicine, University Hospital Essen, Essen, Germany
[4]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[6]University of the Philippines Manila College of Medicine, Ermita Manila, Philippines
[7]Department of Critical Care, Guy's and St Thomas' NHS Foundation Trust, London, UK
[8]Department of Psychiatry, Yale School of Medicine, New Haven, CT, USA
[9]Harvard University, Cambridge, MA, USA
[10]Division of Hematology/Oncology, Department of Pediatric Oncology, Boston Children's Hospital, Boston, MA, USA
[11]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA
[12]Department of Radiology, Emory School of Medicine, Atlanta, GA, USA

which will grow and diminish in value in this context, and place a particular emphasis on the role of clinicians in addressing the safety, ethics, and bias vulnerabilities present with such models. While we emphasize LLMs as particularly transformative tools, we seek to situate them in the broader health informatic landscape, and offer an overview of broad informatics, AI-specific, and LLM-specific skills. Finally, we examine the existing context of medical education, and offer an outline of changes which will be necessary to prepare the healthcare workforce to best take advantage of opportunities afforded by these tools while avoiding their associated pitfalls.

Envisioning the potential of language models in healthcare

Since Google's invention of the Transformer architecture in 2017[4], the LLM field has rapidly advanced in both capability and complexity. At the most fundamental level, the goal of these models is to predict and produce the next token (word or part of word) in a string of text. However, the resulting models demonstrate significant emergent behaviors and the ability to interpret a wide range of inputs to produce complex, nuanced, and contextually-appropriate text. Thus, LLMs have shown robust results in natural language processing applications as a deep learning algorithm for general-purpose language and textual interpretation.

At the clinical level, these models have demonstrated the ability to answer complex, context-specific medical knowledge questions accurately [1, 5, 6], as well as to structure and summarize clinical data [7]. It is not difficult therefore, to envision a future wherein many aspects of day-to-day clinical practice are LLM-facilitated. At the research level, LLMs hold promise to aid in understanding and generating healthcare insights from clinical records and other health-adjacent databases by recognizing, summarizing, translating, predicting and generating text from massive training datasets [5]. Given the rapid expansion of the broad corpus of medical literature, LLMs may play an important role in enabling the generation of up-to-date and patient-relevant guidelines.

Furthermore, the flexibility of the transformer architecture means that combining models across modalities can create a powerful multimodal model capable of leveraging information from several contexts, including radiological images, patient notes, lab tests, and genomics data for example [8–10]. These models have already shown impressive capabilities in the generation of radiology reports as well as prediction of clinical diagnosis and patient outcomes [11–13]. As the increasing digitization of health systems enables the greater linking of Electronic Medical Record (EMR) data, multimodal models may play an increasing role to facilitate greater contextualised predictions.

Understanding the risks of language models in healthcare

However, despite the promise of machine learning in healthcare (MLHC) broadly, there are also significant risks of patient harm. Many models—including those which have been implemented in clinical practice [14, 15]—have been demonstrated to function poorly for minority patient populations. Controversy also exists regarding the "explainability" of such models, and the difficulty in clearly examining the processes which lead to model decisions [16].

There are also risks specific to the LLM architecture and the primacy of text [17]. Firstly, such models have no robust causal models of the world, given that they rely primarily on associative means. LLMs are also prone to so-called "hallucinations", wherein plausible-sounding but incorrect results are generated with apparent confidence [18, 19], including at times the generation and citation of completely fabricated scientific literature.

Further, LLM models are generated based on broad corpora of unselected text which reflect past and present disparities and biases. For instance, pre-trained text embedding models have been shown to offer differential performance on characteristics such as sex and race, predicting that a belligerent white patient is "sent to *hospital*" while a belligerent African American patient is "sent to *prison*" [20].

Section 1: the changing role of clinicians

As stated in the popular maxim of Amara's law, there is a tendency to overestimate the impact of a technology in the short term, and underestimate it in the longer term [22]. Deep learning pioneer Geoffrey Hinton famously opined in 2016 that "We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists" [23]. At the seven year mark, this prediction has yet to come to fruition. Other commentators have predicted the relationship between clinicians and AI to be one of augmentation rather than replacement [24, 25], and experimental evidence has already demonstrated opportunities for collaboration [26]. Similar considerations are likely to apply in the case of LLMs, where achieving potential while minimizing risk will require not only modification of the technology, but also adaptations on the part of clinician users.

Shifts in necessary clinician skills and priorities have always occurred alongside technological changes. For example, the ubiquity and ease of access to bedside reference resources, such as UpToDate or DynaMed, has already begun to reduce the relative value of rote memorization [27, 28]. In its place, however, the ability to rapidly read and summarize literature has grown in prominence with the proliferation of clinical trials and the web search technology. As discussed, the

wide-ranging nature of these models is likely to rearrange the importance of a wide range of clinician skills.

Further, the power of this technology is such that the ideal frame of understanding may shift from seeing models in terms of their usefulness to clinicians, to seeing both clinicians and models in terms of their usefulness to the overall medical system. This concept has been explored in other safety-critical domains such as aviation, regarding human and machine components as cooperative elements of a shared system [29, 30]. While human actors in these domains have advantages over their machine counterparts such as creativity in novel situations, they also have relative deficits such as fatigability and cognitive bias. An ideal system is crafted to ensure complementarity of both sets of capabilities.

Given the rapid pace of change and advancement of this technology, flexibility is likely to be a core virtue on the part of providers. Depending on the rapidity of progress, shifts in the use of these technologies may occur every few years. Therefore, providers must be able and willing to evolve in their practice alongside the evolution of technology, while simultaneously demanding rigorous evaluation of these technologies to determine the robustness and utility of their applications. Clinicians, and their organizations, will be called upon to discern which technologies on the market meet their specific needs, while working with limited resources.

The role of clinicians as repositories for medical knowledge may decrease, as such information becomes increasingly accurately embedded in, and easily accessed from online sources and LLM models. At the same time, the evaluator role of clinicians in medical decision making is going to increase, as they are called upon to evaluate and integrate information offered by models in specific clinical contexts. Clinicians must apply their professional judgment in evaluating model outputs, and use the information as an aid to decision-making rather than a replacement for robust reasoning. Clinicians must also understand the limitations of these technologies with respect to their specific patient population. However, it is important to ensure that these technologies do not lead to an abdication of responsibility, and clinicians are held responsible for their endorsement of decisions facilitated by clinical algorithms.

Similar considerations apply to the generation and application of clinical documentation. AI assistance may lead to less time and energy being spent on the summarization of visits, the generation of documents, and the completion of administrative tasks [31]. But it must be understood that summarization remains a critical element of documentation that has legal and ethical implications, and must not be assumed to be a simple, low-risk space for LLM implementation. Clinical documentation plays an important legal role, and clinicians must accept their legal and ethical responsibility for the comprehensiveness and accuracy of documentation written and signed.

In addition, the empathic role of clinicians in understanding a patient's personality and values will remain essential. Every patient's health and experience of illness is influenced by complicated biological, psychological, and social contexts which cannot be fully reduced through associative summarization. The information returned by LLMs must be carefully contextualized, with the recognition that an "ideal treatment course" may vary substantially based on the idiosyncratic values and preferences of an individual patient. Clinicians must be trusted to act as strong advocates on the part of patients, ensuring that the use of these models remains grounded in foundational principles of medical ethics, and a shared sense of understanding [21, 32, 33]. On the efficiency front, it should be noted that LLMs are able to increase physician engagement with patients only if physicians are not pressured to translate increased efficiency into higher overall output, for example, seeing more patients within a shorter span of time.

Discernment should also be exercised in the study and development of LLMs for clinical use, ensuring that risks are mitigated and ethics are embedded into the models themselves. We must thus seek to train specialized clinician-scientists and leaders who will play an important role in the design, evaluation, and implementation of any LLM-based technologies in healthcare. Given the broad degree of hype, the technology is at risk of being a "solution in search of a problem", used in clinical circumstances for which it is not appropriate or optimal. Thus, judicious development and application of LLMs should be practiced.

Any machine learning models used in healthcare require careful ongoing monitoring due to the challenge of "dataset shift", and the concern that the performance and accuracy of model outputs will decrease over time based on changes in the underlying clinical reality [34]. In addition, the well-known risks related to bias and inequity must be understood, and addressed as foundational challenges rather than secondary afterthoughts [32]. Clinician leaders must work to establish organizational administrative processes in order to satisfy these requirements, and must not abdicate this responsibility to those more distant from patient concerns. Similar concerns apply to the need for development of appropriate regulatory frameworks [35].

Section 2: what all clinicians must know, and what AI specialists should know

As the body of medical knowledge has grown over the past centuries, so too has the demand for specialization. However, medical students do not train exclusively

in one field; they rotate through each specialty, gaining an understanding of each area. The expectation is that clinicians gain an appreciation of the remit of each specialty, an awareness of key emergencies/red flags, and knowledge regarding when to seek specialist advice. A similar dynamic will likely play out with AI, with the dual requirement for developing a baseline of understanding for all clinicians, as well as a more specific body of knowledge for specialists in the field [36].

As the physiological increasingly turns computational, it is easy to mistake the observed for the real, and decision support for dogma. All clinicians should have a baseline set of competencies, including knowledge of (1) forms of bias that can occur in clinical data, (2) uncertainty of predictions, (3) causal inference and confounding. Clinicians must have an understanding of how these systems apply to their individual patients, with particular emphasis on issues of equity and an awareness that such systems do not perform equally for all. For an illustrative example, generalists require only a cursory awareness of the particle spin physics underlying MRI scans, but must have a clear understanding of when an MRI should be ordered for a given patient, how to interpret and explain its results in a clinical context, and when an MRI may be of limited or even misleading utility.

In addition to these baseline competencies, we must seek to train AI Specialist-Clinicians with a combined understanding of medicine and computer science, able to play important roles in the development and implementation of these models [37]. Such clinicians will require a much deeper understanding of fundamental model architecture, and the capabilities and limitations of a given approach. In addition to their work in model development and monitoring, we envision such clinicians

Table 1 Selected competencies for General and AI specialist-clinicians

	General Clinician	AI Specialist-Clinician
Health Informatics Competencies	- Knowledge of data stewardship and patient privacy concerns. - Ability to work with electronic medical records in recording and accessing patient information. - Ability to engage with telemedicine systems. - Awareness of the limitations of health data systems with respect to completeness and representativeness of data. - Ability to adapt to and work with novel informatics interfaces and computer systems. - Ability to use basic clinical decision support systems.	- Detailed understanding of clinical workflows, with the ability to appropriately design models to support given clinical use cases. - Understanding of the social, economic, and political context of AI at the level of health research, health system structure, and health technology regulation. - Organizational management skills to guide informatics implementation projects and workflows. - Proficiency in data management skills, such as data cleaning and quality assurance - Understanding of and ability to align work with common data interoperability standards.
Artificial Intelligence Competencies	- Understanding the applicability of a given AI technology in specific clinical contexts. - Interpreting and explaining the output of a given AI model (or, in the case of unexplainable models, evaluating the empirical validation process of a given model). - Knowledge of the limitations of a given AI model, with a particular emphasis on fairness and bias as well as differential model performance. - Understanding the importance of human oversight and the limitations and failure cases of AI systems. - Ability to recognize and mitigate AI failures as they arise in a clinical workflow.	- Detailed knowledge of model architecture, with assessments of the appropriateness of a specific technology for a given clinical task. - Evaluating the performance and robustness of an AI model for a specific clinical problem. - Evidence-based evaluation of AI-based tools, including trial design, implementation, and continuous monitoring. - Generating and curating datasets for the purpose of - Identifying limitations and biases in performance of algorithms towards marginalized groups and fine-tuning model performances by curating more representative datasets. - Ability to interpret and communicate model performance metrics to non-technical stakeholders. - Awareness of the legal and regulatory landscape for AI in healthcare, including liability concerns and approval processes.
Generative AI / LLM Competencies	- Baseline awareness of the inputs and architectures of LLMs. - Skills in offering both initial and follow-up queries to LLM systems as appropriate in a given clinical context. - Skills in prompt engineering, with awareness of the context-specificity and stochasticity of LLM outputs. - Understanding of the "hallucination" phenomenon, and ability to be appropriately skeptical and verify LLM outputs where necessary. - Integrating information from a diverse range of sources (including LLM summaries or differential diagnostic predictions alongside traditional info such as patient demographics, clinical presentation, and investigation results) in the context of a patient-centered clinical encounter.	- Collaboration with colleagues from other medical specialties to identify opportunity and limitations for further development of LLMs within clinical contexts. - Ability to generate and incorporate human feedback and clinician / patient preference information in fine-tuning models. - Detailed generation and evaluation of prompts, alongside sensitivity analysis for the impact of subtle prompt fluctuations on outputs - Understanding of cutting-edge technical developments in generative AI (such as retrieval-augmented generation (RAG) or long context window techniques). - Ability to design and implement human evaluation studies to assess clinical impacts of LLM-generated content. - Understanding of the ethical and legal implications of using generative AI, including with respect to intellectual property, liability, and privacy.

being able to offer specific consultations on AI-related questions. An instructive analogy here would be the role of a radiologist in helping to guide appropriate imaging modality selection, and discussing the implications of unclear or unexpected results.

Section 3: situating LLM skills in the broader context

The skills necessary to take advantage of LLMs do not exist in isolation. Rather, they must be understood in the context of a long history of foundational work on competencies in health informatics [38] as well as more recent work on AI broadly [36]. Rather than bypassing these foundations and "reinventing the wheel", LLM-specific education must be complementary. As summarized in Table 1, both the general clinician and the AI specialist clinician require skills in all three domains in order to find success.

Whether in sourcing the data for their inputs, or communicating their outputs to clinicians, many modern LLM-based systems are being constructed in tandem with EMRs [39] and their underlying databases. While there is some prospect of utilizing LLMs themselves to improve the quality and structure of EMR data [39, 40], many of the existing challenges with medical data persist and indeed may only grow in importance alongside the increasing power of medical AI. Ability to efficiently and effectively use these systems while safeguarding patient privacy and interests is important to every physician. Ai specialist leaders must be able to go further, and effectively lead the organizational transitions necessary for health data to contribute to care-enhancing systems in real time.

With AI systems moving from merely providing information to providing recommendations and more specific judgments, the skills required of practitioners grow in specificity [36]. This entails knowledge of the specific strengths and weaknesses of such systems, with a particular emphasis upon the cases where they may underperform or outright fail. While the general clinician must learn to appropriately engage with information and recognize such failures, the AI specialist clinician must be able to go further, designing and evaluating systems to mitigate failures in the first place. AI specialists must also be able to navigate the complicated social, legal, and ethical implications of these technologies, and translate between technical and non-technical stakeholders in both development and implementation.

Engagement with generative AI — and LLMs specifically — requires these skills to be nuanced and focused upon the particular strengths and limitations of this nascent technology. The novel skill of "prompt engineering" (that is, designing text inputs to achieve the appropriate outputs) is necessary both at the level of generalists and AI specialists [41]. While the former must be able to effectively prompt in the context of a clinical workflow, the latter must be able to systematically analyze and optimize prompting techniques (which, as recent research has shown [42], can vastly improve model performance). LLMs are also unique in their ability to be confidently persuasive while "hallucinating" [43], necessitating a degree of appropriate skepticism from clinicians, and the ability to verify information given. With the field so rapidly evolving, it is important for AI specialist clinicians to keep abreast of the technical detail of new generative AI research, and connect it effectively to medical practice.

Section 4: required adaptations of medical education

Despite the increasing interest in artificial intelligence in medicine, existing approaches to medical education on the subject are inconsistently offered, and have highly variable content between centers [44]. LLM-specific education is relatively nonexistent at this time. We believe that the tremendous rate of advancement in these technologies necessitates significant ongoing investment in addressing medical education challenges in this regard. Students entering medical school in 2024 will enter independent practice in the early- to mid-2030s, to say nothing of the training that existing clinicians will require. Training must be forward-looking, and early steps must be taken with urgency to ensure that the medical system is up to the task [45].

First is the question of "what". There is little agreement regarding what core competencies must be taught to medical students regarding this topic, with major differences between institutions such as Harvard University, and the University of Toronto [36, 46]. Broad, multistakeholder efforts are required to establish shared competency frameworks and enable collaboration to develop collective resources [47]. This may take the form of a specific analogue to existing efforts such as the development of the CanMEDS framework, outlining multiple domains of broad clinician competency [48]. In addition, there must be a core recognition (as we have outlined above) of the difference between what *all* must know and what must be done to train clinical AI specialists [36, 49].

Second is the question of "who". The interdisciplinary nature of AI technologies requires the insight and expertise not only of clinicians, medical educators, and their existing primary collaborators in pre-clinical departments, such as anatomy and physiology. On the technical side, expertise is needed from computer scientists, engineers, cybersecurity experts, and bioinformaticians among others. Further, in order to understand the context and risks of these models, input is required from ethicists, sociologists, medical anthropologists, and others. The intensely complicated and multi-disciplinary

nature of this technology will necessitate the creation of new professional relationships and organizations cutting across these divides. This will require upstream adaptation of existing hierarchies and recruitment pathways in medical schools, in order to create a robust and responsive faculty base.

Third is the question of "how". Given how little consensus exists regarding what must be taught [50], it is unsurprising that there is little consensus regarding the optimal methods of AI education. We believe that the recent progress has significant implications not only for AI-specific education, but also for medical education more broadly. Progress in this field must prompt a more radical reconceptualization of medical education broadly. The relative ease with which these models are able to pass the USMLE examination, for example, may call into question the structure of this examination process, and the significant portion of study time which is spent on the rote memorization of facts. With wide bodies of factual information at the fingertips of clinicians, we posit that there must be a broad shift from an emphasis on *knowledge* to an emphasis on *skills*. It is encouraging that over the last year, the grading of the USMLE exam has moved to Pass/Fail in place of a numeric score [51].

Concurrently, we advocate for a transition away from the more traditional medical education model to a pathway-driven model in which learners can leave medical school with a deeper area of expertise in a chosen domain (e.g. medical computing, public policy/social medicine, clinical research, etc.). In this way, medical education better mirrors other professional schools such as business, law, and engineering and would provide the necessary time to specialize in important contemporary topics, such as AI and LLMs. Similarly, as students use their foundational medical training to determine their field of practice, they will also have the opportunity to hone non-clinical skills and knowledge and discover ways to impact medicine on a larger scale.

Attention in this regard must not be solely focused at the undergraduate and postgraduate medical education levels. The majority of users of LLM based technology, if it is implemented in healthcare within the next decade, will be existing clinicians who have received little to no specific training in the topic. Efforts must be made on the continuing medical education front to prepare clinicians to effectively use these technologies. Of equal importance will be research and examination of existing healthcare processes on the front of AI technology, and to ensure safe and effective implementation.

It is also crucial to acknowledge the contributions of established educational theory in this process. Constructivist theory, in particular, aligns well with the challenges of teaching AI skills in a rapidly evolving field. This theory emphasizes that learners actively construct knowledge through experiences and interactions with their environment, rather than passively receiving information [52]. In the context of AI education for medical professionals, constructivist approaches would involve experiential, active learning within clinical care settings. This is especially relevant as both learners (medical students and residents) and teachers (supervising physicians) may be simultaneously developing their AI competencies. Problem-based learning, case studies, and hands-on projects involving real-world AI applications in healthcare could be effective strategies rooted in constructivist theory. This approach not only helps in skill development but also cultivates critical thinking and adaptability – crucial attributes for navigating the dynamic landscape of AI in medicine.

Ultimately, changes to the how, who and what of medical education should also reflect on grading and testing practices. As LLMs have repetitively been shown to easily pass standardized tests, such as USMLE, this raises the urgent question if a transition away from knowledge to understanding, reasoning and critical thinking is needed [53].

Conclusion

Recent advances in LLM technology have been striking in both their magnitude and pace, raising the prospect for a future where such technologies are deeply integrated into clinical practice. These developments bring the opportunity for improving the quality of care through rapid summarization and presentation of medical knowledge, as well as significantly reducing administrative burden. At the same time, there are significant risks associated with these models, particularly in relation to datasets embedding existing societal biases, and the specific tendency of LLMs toward overconfident "hallucination" of answers.

Nonetheless, these systems are primed to have a wide ranging series of impacts on healthcare, and all clinicians must be effectively trained to take advantage of these opportunities while countering the associated risks. This task will require a collective undertaking toward understanding the "what", "who", and "how" of AI-related medical education, in order both to establish a baseline level of competence among clinicians broadly, and to facilitate the training of clinicians with a deeper level of AI-specific expertise. Medical education for students and practicing clinicians alike must adapt with rapidity and dynamism matching the pace of technological change.

Abbreviations

LLM	Large Language Model
MLHC	Machine Learning for Healthcare
USMLE	United States Medical Licensing Examinations
AI	Artificial Intelligence
EMR	Electronic Medical Record

Acknowledgements
We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

Author contributions
Initial conceptions and design: LGM, FYCN, LAC, CMS. Drafting of the paper: All Authors Critical revision of the paper for important intellectual content: All Authors.

Funding
CM Sauer is supported by the German Research Foundation funded UMEA Clinician Scientist Program, University Hospital Essen, grant number FU356/12 – 2.
J.G. declares support from US National Science Foundation (grant number 1928481) from the Division of Electrical, Communication & Cyber Systems, RSNA Health Disparities grant (#EIHD2204) and NIH (NIBIB) MIDRC grant under contracts 75N92020C00008 and 75N920.
L.A.C. is funded by the National Institute of Health through NIBIB R01 EB017205.
Open Access funding enabled and organized by Projekt DEAL.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 February 2024 / Accepted: 18 September 2024
Published online: 07 October 2024

References

1. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2:e0000198.
2. Yang X, PourNejatian N, Shin HC et al. GatorTron: A Large Language Model for Clinical Natural Language Processing. 2022; : 2022.02.27.22271257.
3. Google AI, Blog. Our latest health AI research updates. Google. 2023; published online March 14. https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/ (accessed March 19, 2023).
4. Vaswani A, Shazeer N, Parmar N et al. Attention Is All You Need. arXiv.org. 2017; published online June 12. https://arxiv.org/abs/1706.03762v5 (accessed March 19, 2023).
5. Singhal K, Azizi S, Tu T et al. Large Language Models Encode Clinical Knowledge. arXiv.org. 2022; published online Dec 26. https://arxiv.org/abs/2212.13138v1 (accessed March 19, 2023).
6. OpenAI. GPT-4 Technical Report. arXiv.org. 2023; published online March 15. https://arxiv.org/abs/2303.08774v2 (accessed March 19, 2023).
7. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. Large Language Models are Few-Shot Clinical Information Extractors.
8. Meskó B. The impact of Multimodal large Language models on Health Care's future. J Med Internet Res. 2023;25:e52865.
9. Zhang S, Xu Y, Usuyama N et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. 2024; published online Jan 16. https://doi.org/10.48550/arXiv.2303.00915
10. Multimodal Learning With Transformers. A Survey. https://www.computer.org/csdl/journal/tp/2023/10/10123038/1N3MioQlClW (accessed April 2, 2024).
11. Tu T, Azizi S, Driess D, et al. Towards Generalist Biomedical AI NEJM AI. 2024;1:AIoa2300138.
12. Khader F, Kather JN, Müller-Franzes G, et al. Medical transformer for multimodal survival prediction in intensive care: integration of imaging and non-imaging data. Sci Rep. 2023;13:10666.
13. Zhou H-Y, Yu Y, Wang C, et al. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. Nat Biomed Eng. 2023;7:743–55.
14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447–53.
15. Huang J, Galal G, Etemadi M, Vaidyanathan M. Evaluation and mitigation of racial Bias in Clinical Machine Learning models: scoping review. JMIR Med Inf. 2022;10:e36388.
16. McCoy LG, Brenna CTA, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. J Clin Epidemiol. 2022;142:252–7.
17. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, 2021: 610–23.
18. Ji Z, Lee N, Frieske R et al. Survey of Hallucination in Natural Language Generation. arXiv.org. 2022; published online Feb 8. https://doi.org/10.1145/3571730
19. Manakul P, Liusie A, Gales MJF, SelfCheckGPT. Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv.org. 2023; published online March 15. https://arxiv.org/abs/2303.08896v1 (accessed March 19, 2023).
20. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In: Proceedings of the ACM Conference on Health, Inference, and Learning. Toronto, Ontario, Canada: Association for Computing Machinery, 2020: 110–20.
21. Harrer S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. eBioMedicine. 2023;90:104512.
22. Amara's law. Wiktionary. 2020; published online July 13. https://en.wiktionary.org/w/index.php?title=Amara%27s_law&oldid=59741401 (accessed March 19, 2023).
23. Geoff Hinton: On Radiology. 2016 https://www.youtube.com/watch?v=2HMPRXstSvQ (accessed March 19, 2023).
24. Ahuja AS. The impact of artificial intelligence in medicine on the future role of the physician. PeerJ. 2019;7:e7702.
25. Will AI, Eventually Replace Doctors? Kellogg Insight. 2023; published online Feb 1. https://insight.kellogg.northwestern.edu/article/will-ai-replace-doctors (accessed March 19, 2023).
26. Reverberi C, Rigon T, Solari A, Hassan C, Cherubini P, Cherubini A. Experimental evidence of effective human–AI collaboration in medical decision-making. Sci Rep. 2022;12:14952.
27. Baxter SL, Lander L, Clay B, et al. Comparing the Use of DynaMed and UpToDate by Physician trainees in clinical Decision-Making: a randomized crossover trial. Appl Clin Inf. 2022;13:139–47.
28. Wartman SA, Combs CD. Reimagining Medical Education in the age of AI. AMA J Ethics. 2019;21:E146–152.
29. Hoc J-M. From human – machine interaction to human – machine cooperation. Ergonomics. 2000;43:833–43.
30. McCoy LG, Burkell J, Card D et al. On Meaningful Human Control in High-Stakes Machine-Human Partnerships. 2019.
31. van Buchem MM, Boosman H, Bauer MP, Kant IMJ, Cammel SA, Steyerberg EW. The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med. 2021;4:57.
32. McCoy LG, Banja JD, Ghassemi M, Celi LA. Ensuring machine learning for healthcare works for all. BMJ Health Care Inf 2020; 27.
33. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25:1337–40.
34. Finlayson SG, Subbaswamy A, Singh K, et al. The clinician and dataset shift in Artificial Intelligence. N Engl J Med. 2021;385:283–6.
35. Gichoya JW, McCoy LG, Celi LA, Ghassemi M. Equity in essence: a call for operationalising fairness in machine learning for healthcare. BMJ Health Care Inf. 2021;28:e100289.
36. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? Npj Digit Med. 2020;3:1–3.
37. Cussat-Blanc S, Castets-Renard C, Monsarrat P. Doctors in Medical Data sciences: a New Curriculum. Int J Environ Res Public Health. 2022;20:675.

38. Jidkov L, Alexander M, Bark P, et al. Health informatics competencies in postgraduate medical education and training in the UK: a mixed methods study. BMJ Open. 2019;9:e025460.
39. Nashwan AJ, AbuJaber AA. Harnessing the power of large Language models (LLMs) for Electronic Health Records (EHRs) optimization. Cureus. 2023;15:e42634.
40. Guevara M, Chen S, Thomas S, et al. Large language models to identify social determinants of health in electronic health records. NPJ Digit Med. 2024;7:6.
41. Meskó B. Prompt Engineering as an important emerging skill for medical professionals: Tutorial. J Med Internet Res. 2023;25:e50638.
42. Nori H, Lee YT, Zhang S et al. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. 2023; published online Nov 27. https://doi.org/10.48550/arXiv.2311.16452
43. Huang L, Yu W, Ma W et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 2023; published online Nov 9. https://doi.org/10.48550/arXiv.2311.05232
44. Civaner MM, Uncu Y, Bulut F, Chalil EG, Tatli A. Artificial intelligence in medical education: a cross-sectional needs assessment. BMC Med Educ. 2022;22:772.
45. Lomis K, Jeffries P, Palatta A, et al. Artificial Intelligence for Health Professions Educators. NAM Perspect. 2021;2021. https://doi.org/10.31478/202109a.
46. Law M, Veinot P, Campbell J, Craig M, Mylopoulos M. Computing for Medicine: can we prepare medical students for the future? Acad Med. 2019;94:353.
47. Russell RG, Lovett Novak L, Patel M, et al. Competencies for the Use of Artificial Intelligence-based tools by Health Care professionals. Acad Med. 2023;98:348–56.
48. Frank JR, Danoff D. The CanMEDS initiative: implementing an outcomes-based framework of physician competencies. Med Teach. 2007;29:642–7.
49. Wiljer D, Hakim Z. Developing an Artificial intelligence–enabled Health Care Practice: Rewiring Health Care professions for Better Care. J Med Imaging Radiation Sci. 2019;50:S8–14.
50. Ngo B, Nguyen D, vanSonnenberg E. The cases for and against Artificial Intelligence in the Medical School Curriculum. Radiol Artif Intell. 2022;4:e220074.
51. USMLE Step 1 Transition to Pass/Fail Only Score Reporting | USMLE. https://www.usmle.org/usmle-step-1-transition-passfail-only-score-reporting (accessed March 21, 2023).
52. Dennick R. Constructivism: reflections on twenty five years teaching the constructivist approach in medical education. Int J Med Educ. 2016;7:200–5.
53. Abbas A, Rehman MS, Rehman SS. Comparing the performance of Popular large Language models on the National Board of Medical Examiners Sample Questions. Cureus. 2023;16:e55991.

Publisher's note