2024, Volume 56, Issue 9, 534-540, e-ISSN 1938-3800

Family Medicine

SPECIAL ARTICLE

Generative Artificial Intelligence and Large Language Models in Primary Care Medical Education

Daniel J. Parente, MD, PhD

AUTHOR AFFILIATION:

Department of Family Medicine and Community Health, University of Kansas Medical Center, Kansas City, KS

CORRESPONDING AUTHOR:

Daniel J. Parente, Department of Family Medicine and Community Health, University of Kansas Medical Center, Kansas City, KS, dparente@kumc.edu

HOW TO CITE: Parente DJ. Generative Artificial Intelligence and Large Language Models in Primary Care Medical Education. *Fam Med.* 2024;56(9):534-540. doi: 10.22454/FamMed.2024.775525

PUBLISHED: 8 August 2024

KEYWORDS: artificial intelligence, curriculum development, graduate medical education, information technology, undergraduate medical education

ABSTRACT

Generative artificial intelligence and large language models are the continuation of a technological revolution in information processing that began with the invention of the transistor in 1947. These technologies, driven by transformer architectures for artificial neural networks, are poised to broadly influence society. It is already apparent that these technologies will be adapted to drive innovation in education. Medical education is a high-risk activity: Information that is incorrectly taught to a student may go unrecognized for years until a relevant clinical situation appears in which that error can lead to patient harm. In this article, I discuss the principal limitations to the use of generative artificial intelligence in medical education—hallucination, bias, cost, and security—and suggest some approaches to confronting these problems. Additionally, I identify the potential applications of generative artificial intelligence to medical education, including personalized instruction, simulation, feedback, evaluation, augmentation of qualitative research, and performance of critical assessment of the existing scientific literature.

THE AGE OF ARTIFICIAL INTELLIGENCE

Educational methodology has repeatedly adapted to technological change.[1] Technological revolutions are an enduring feature of human societies and have been defined as "dramatic change brought about relatively quickly by the introduction of some new technology"[2] with several examples of technological revolutions given, including the development of agriculture 11,000 years ago, the invention of movable type printing in 1448, and the development of atomic physics and quantum mechanics in the middle of the 20th century.[2]

One of the more recent adaptations in educational methodology has been driven by the rise of computers. Computers were made possible by the revolution in quantum mechanics in the mid-20th century, which directly led to the creation of transistors and other semiconductor devices. Transistors now form the basis of all computing technology; more than 13 sextillion transistors have been manufactured.[3] Transistor technology was first developed in December 1947 at Bell Labs in Murray Hill, New Jersey. At that site, Walter Brattain, John Bardeen, and William Shockley applied an electric current to one of two closely spaced gold foil plates in contact with purified germanium.[4] From the second gold plate, out came an amplified version of this current, thereby creating the first transistor.

There is a gap, however, between when a new technology is developed and when its potential has been realized. On the day that transistors were invented in New Jersey, the entire world had *already* fundamentally changed even though only three people knew about it. What was technologically *possible* had changed, even though it would take decades for the flourishing and interconnectivity of computers to become actualized. Because of transistor technology, the set of activities humanity could carry out expanded dramatically. The technological rules of the game had changed.

By the preceding definition, a new technological revolution, built on the successes of the transistor revolution is likely in progress: the artificial intelligence (AI) revolution. Although AI research has existed as an organized domain of study since 1956,[5] only recently has a crucial breakthrough ushered in rapid progress. In 2017, Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin published "Attention Is All You Need," describing the transformer architecture for artificial neural networks (ANNs).[6] ANN models can be trained on a dataset,[7] which effectively accomplishes a very complicated, nonlinear fit to the data. When using ANNs to model human languages, the model must be able to parse out relationships between words in a sentence.[6] For example, in the sentence "The man saw the lion as he roared," the model must encode that the "he" who roared is likely "the lion" rather than "the man." Prior solutions to this problem involved using so-called "recurrent neural networks," but such networks suffered from limited parallelism and took a large amount of time to train.[6] The transformer architecture eliminated recurrence from ANNs—without sacrificing the ability to encode sematic relationships between words—thereby allowing marked gains in parallelism and training time, enabling very large data sets to be used for training.[6]

Transformers' neural network architectures have rapidly given rise to a wide range of algorithms in the class of "generative artificial intelligence." Among these algorithms, large language models (LLMs)—such as OpenAI's ChatGPT, Meta's Llama models, and Google's Gemini models—have been driving innovation across all areas of society, including medicine and education.[6,8–15] LLMs encode a statistical understanding of human languages within a neural network by analysis of vast volumes of training text enabled by transformer neural networks.[16] LLMs are trained to perform next-word prediction.[16] Once trained, LLMs can be used to generate text that is convincingly humanlike.[16] Importantly, LLMs appear to be flexible and can be instructed to conduct many tasks, such as text generation, summarization, logical inference, and computer programming.[6,8–12,17–19]

Technological revolutions do not limit their influence to a single domain; they affect society widely.[2] As clinician educators, it is appropriate to consider how the AI technological revolution will impact medical education. Just as medical education adapted to the advent of the transistor and computing by incorporating digital technologies into medical education, we must rapidly adapt to the advent of generative AI and LLMs.

There is risk here. Risks can be categorized in terms of their impact (trivial to severe) and latency (immediate to years later). Errors in medical education have the possibility of creating severe and long latency risks, such as an error that goes unrecognized for years and leads to the death of a patient. (High severity, long latency risks are not unique to medicine and can occur in other settings; for example, a mistake in the software controlling a traffic light might lead to an accident under conditions that occur infrequently.) Especially as society struggles with artificially intelligent systems that are at the borderline of competence as computerized educators, we must remain vigilant and insist that applications of generative AI and LLMs be safe and validated before widespread use.

Here, I discuss issues related to the implementation of LLMs in medical education with an emphasis on primary care and suggest pathways for safe and effective use of generative AI in this context.

PROBLEMS WITH GENERATIVE ARTIFICIAL INTELLIGENCE

Widespread deployment of LLMs requires them to be accurate, bias-free, affordable, and secure. Challenges in all these domains exist for generative AI.

Hallucination

For LLMs to be used in primary care and medical education, they must supply correct information. Because LLMs have billions of parameters, they can encode a large amount of information about the world within the models directly. For GPT-3.5, with 175 billion parameters,[9] the model likely encodes about 85 gigabytes of data, which is approximately the uncompressed equivalent of all text in Wikipedia.*

Clearly these models can contain much knowledge about the world—but not an unlimited amount. When asked questions for which the model has no knowledge, many LLMs will, unfortunately, produce replies that appear to be authoritative but are not based in reality.[20–23] This has been termed "hallucination" (see Maleki et al[24] for a brief history of the term and analysis of some various alternatives terms). Hallucination is one of the principal limitations for safe use of LLMs. The risk of hallucination is that a student will ask a specific technical question to which the LLM will supply a confident but wrong answer, thereby causing mislearning. Reducing hallucination is an open research topic within computer science, and further progress is expected imminently. Medical educators must insist that LLM-based medical education tools undergo validation to quantify the degree to which hallucination impacts performance.

Bias

Historical and structural factors have negatively influenced health care outcomes for disadvantaged and minority populations.[25,26] A large volume of scholarship discusses how medical education must guard against incorporating structural biases into medical education curricula.[27–33] Indeed, repeated calls have been made for the development and implementation of structural bias curricula both across medical education and within family medicine clerkships.[29–33] AI, machine learning, and algorithmic clinical prediction algorithms can produce biased results.[34–38] (Note that bias is not necessarily an unavoidable result, because neural network deep learning approaches also have been used to reduce bias.)[39] LLMs, too, run the risk of perpetuating biases present in their training data, although measures are being taken by developers to mitigate them.[16]

Technology-related and algorithmic biases may be subtle or become insidiously entrenched. For example, recent work has highlighted the risk of racial bias in even apparently benign clinical technologies such as pulse oximetry and the estimating equations for glomerular filtration rate.[40,41] Moreover, some historically accumulated medical evidence is invalid due to improper study designs that neglect the diversity of the population.[42] For example, study designs that over- or underrepresent racial, ethnic, socioeconomic, or sexual minority groups risk poor generalizability to the rest of the population. Extensive efforts are underway to compile datasets that reflect the diversity of the entire population.[42,43]

Active research is ongoing on algorithmic ways of mitigating bias in LLMs.[38] These techniques operate both during the training of models and after models have been trained, and are in use. An example of training time bias mitigation is counterfactual data augmentation (CDA). If CDA were used to mitigate gender bias, gendered words (eg, "he" and "she") can be sometimes swapped in the training dataset to rebalance the dataset.[38,44,45] Other bias-mitigation strategies also have been explored.[38] Before LLM-based systems are trusted for medical education—even those including algorithmic components to mitigate bias—they will require validation by human experts who have training in recognizing structural and systemic biases to avoid unintentionally propagating biases to medical learners via AI.

Cost

LLMs are computationally intensive mathematical models that require computing hardware with many parallel processing units and many gigabytes of memory. Because of this need, professional-grade graphical processing units (GPUs) have been required. This has meant that most LLMs have been run on servers using a cloud computing framework. For example, OpenAI maintains access to the GPT-3.5 and GPT-4 models and charges for access based on the total number of tokens sent and received. A "token" is about four characters or, on average, 75% of a word. For the highest performance model, GPT-4, the current price is 3 cents per thousand input tokens and 6 cents per thousand output tokens. For multiturn conversations, each turn in the conversation involves processing all previously sent and received tokens. Consequently, these costs can quickly add up. The primary care medical education community should be cautious to select LLMs that can be used in an economically sustainable way.

The LLM research and enthusiast community has been experimenting with methods of running LLMs on local (not server) hardware using consumer-grade graphic processing units (GPUs; see, eg, Dettmers et al[46], Frantar et al[47], Gerganov[48]). Because Meta chose to make its Llama models public, the early work of this community was primarily focused on the Llama models.[49,50] Various Llama models have been released, using between 7 billion and 70 billion parameters. Larger models, in general, yield better performance.[46] Attempting to run these models locally would require between 28 and 280 gigabytes of video random access memory (RAM) in a local graphics card. This amount is unavailable with the most powerful consumer-grade GPUs, which invariably have less than 24 gigabytes of RAM.

Groundbreaking experimental and theoretical work on quantizing the parameters from a 16-bit number (ranging from 0 to 65536) into a 4-bit number (ranging from 0 to 15) has been able to fit medium-sized models—those with 33 billion parameters—into 24 gigabytes of video RAM.[46–48] This renders them suitable for execution on local hardware. By eliminating the need for cloud computing, the costs are then primarily up-front costs of purchasing a suitable GPU and host computer system for the model. This enables LLMs to be used in ways that would be cost-prohibitive if attempting to use cloud computing-based models (eg, making extensive use of multiturn conversations). At my institution, we were able to assemble the necessary components for less than $3,000, which is well within the budgets of academic medical centers and small research projects.

Local LLMs have trade-offs in performance with closed-source cloud models such as OpenAI's GPT-4 Turbo, Anthropic's Claude 3 Opus, and Google's Gemini 1.5 Pro, which still provide superior performance.[51] At least one open-source, local LLM (Meta's recently released Llama 3 70 billion parameter model) is, however, performance-competitive with these models.[51] The primary care medical education community should carefully consider the balance between cost, security, and performance when choosing between local and cloud-based LLMs during the development of educational tools.

Security

Data security is a key concern in medical education. Medical learners interact with patient data, which is protected under the Health Insurance Portability and Accountability Act (HIPAA). Likewise, educational records are protected by the Family Educational Rights and Privacy Act (FERPA). These statutes impose regulatory burdens on clinicians and educators to protect data appropriately. When cloud computing is used, data must be transmitted outside of an academic medical center to a third party (eg, OpenAI or Google). Adequate provisions for data security under HIPAA and FERPA must be ensured for this transmission to be allowable. Data security may be better accomplished by performing all computations on premises using local LLM strategies (as described previously) to minimize the need to transmit protected information outside of the host medical center. Security-related decisions should, however, be taken in consultation with local institutional experts.

Academic Integrity and Copyright

Discussion is ongoing within society, and within the educational community specifically, as to the appropriate role of LLMs when used for scholarly activities.[52,53] Moreover, there are outstanding legal questions related to the use of copyrighted material during the training and use of LLMs (exemplified by the ongoing lawsuit between The New York Times Company and OpenAI).[54] As end users of these technologies, medical educators would benefit from remaining aware of the changing legal and regulatory landscape.

POSSIBILITIES OF GENERATIVE ARTIFICIAL INTELLIGENCE IN MEDICAL EDUCATION

Despite these problems, primary care medical educators have an imperative to prepare for the use of generative AI and LLMs in medical student and resident education. This technology has numerous future applications.

Personalized Instruction

Most salient is the possibility of delivering personalized instruction. Current didactic approaches to medical education require an instructor to deliver content that is understood by most learners. These strategies must, unavoidably, yield suboptimal outcomes both for learners who are more advanced and for those who are struggling with the material. If LLMs could be validated to provide information that is known to be accurate, free of bias and hallucination—at least as good as human instructors—then students could interact with a personalized virtual tutor and could learn at a pace suitable to their current level.

Students also are likely to differentially struggle with certain aspects of the material in unpredictable ways. Having a personalized instructor to ask specific questions about areas of confusion would allow students to spend their time more effectively by focusing on areas in which they are confused and by not rehashing material they have already mastered. I call for research on whether this technology will improve not only learner satisfaction (eg, Kirkpatrick Level 1), but also whether it will facilitate knowledge transfer (Level 2), thereby leading to changed learner behavior (Level 3) and patient outcomes (Level 4).

Simulation

Simulation can help students practice in a safe environment, which may be particularly relevant for high-acuity situations such as in critical care.[55] Simulation is nevertheless expensive to administer, and the cost-effectiveness of simulation as traditionally implemented has been questioned.[56] Simulation has both fixed (start-up) costs and variable (ongoing) costs.[56] Fixed costs include faculty salary/time to develop clinical scenarios, purchase of equipment (eg, mannequins), and investment in buildings and facilities in which to conduct simulation.[56] Variable costs are principally driven by (a) the personnel costs of staff administering the simulation, and (b) facility charges.[56]

If appropriately validated with respect to hallucination and bias, LLMs could reduce variable costs by acting as virtual standardized patients or virtual simulation administrators for hundreds of students simultaneously at a bare fraction of the personnel costs required to run a simulation on a given day. Moreover, simulations may also reduce fixed costs by partially automating the process of faculty physicians constructing new simulation scenarios, allowing them to refine a rough draft scenario, quickly sketched out by the LLM on demand, rather than needing to start drafting a new clinical scenario entirely from scratch. Human-AI cocreation tools already are being built into creative tools (eg, for software development) and can improve efficiency.[57] I call on the primary care research community to develop pilot LLM-based simulations, validate them with respect to hallucination and bias, and subject them to economic cost-benefit analysis to determine their performance compared to (a) traditional simulation, and (b) other teaching methods (eg, didactic education).

Feedback and Evaluation

Narrative assessment of clinical learners by faculty physicians serves several purposes. First, it provides students guidance on areas of strength and deficiency, allowing adaptation and improvement along their learning trajectory. Second, narrative feedback is important for evaluation both within a clerkship and for standardized evaluations that support residency applications. Quality narrative feedback, however, is difficult to achieve,[58–60] and many interventions seeking to improve feedback quality do not yield improvement, lack sustainability, are not cost effective, or are not generalizable.[61–63] Moreover, faculty physicians may not have time to write constructive, narrative feedback for students because of care and compensation models that emphasize clinical productivity to the detriment of medical education.[64,65] Furthermore, not surrendering evaluative authority to an algorithm incapable of understanding the full context of a student's circumstances and performance will be important. Humans, therefore, ought to be kept in the loop at all times as a general principle.

Because generation of humanlike text is the core competency of LLMs, developers should determine how to form partnership between faculty clinicians and LLMs to allow for higher-quality written feedback that is more constructive and targeted. Validation studies will be necessary to ensure that AI-augmented systems are acceptable to both faculty and students, and do not result in an increase in bias or a decrease in the quality of narrative feedback.

Qualitative Medical Education Research

Many lessons are to be learned from qualitative research projects that analyze free-form responses of learners in survey or interview format.[66] Many such responses must undergo a coding process, which is laborious and time-intensive for the coders.[67] Because LLMs can operate on large volumes of text quickly, the medical education qualitative research community should determine how best to semiautomate, or even fully automate, the initial steps in the coding process. Research progress in this area will make large corpora of qualitative text more readily analyzable, thus accelerating research in primary care medical education. Progress toward this goal is already being made in domains outside medical education.[68,69] Because of the hallucination problem, an initial focus on coding tasks that emphasize a coding schema that is strongly backed by existing theory is likely the most fruitful area. This suggestion is based on observations that LLMs seem to perform better when given clear and detailed instructions (eg, "summarize this text") rather than when directed to do tasks that require judgment and reasoning (eg, "decide whether this is important"),[70] such as discovering themes that have not been previously placed in a theoretical context.

Critical Assessment of Scientific Literature

The biomedical literature is vast. PubMed has 36 million citations, with 1 million added in 2022. Even restricting focus to just scientific journals primarily on primary care or family medicine this amounts to at least 5,000 articles yearly. Both faculty clinicians and medical learners have limited time in which to prioritize which of these articles require review. LLMs can accelerate the surveillance and review of the academic literature, including as it relates to primary care and medical education.[70] I invite the primary care medical education community to work together to find new ways to use LLMs safely and effectively while accelerating critical assessment of the primary care medical education literature.

https://doi.org/10.22454/FamMed.2024.775525

EMBRACING ARTIFICIAL INTELLIGENCE FOR MEDICAL EDUCATION

The AI era has arrived. The world has already changed, although society is still grappling with the full implications these revolutionary technologies will have on our world. As clinicians and educators, we must not fail to use these technologies to enhance medical education and ultimately human health. We must still be always mindful of the pitfalls of these innovative technologies. The problems of hallucinations, bias, cost, and security must be considered when implementing safe and effective tools using generative AI in medical education. Validation will be needed to ensure accuracy of these tools and to minimize the risk of unintentional, systemic bias being transmitted to medical learners.

I recommend prioritizing research into how LLMs—and generative AI, more generally—best enable medical educators to be more effective by improving personalized instruction, simulation, feedback, and evaluation. Research into how LLM-based tools can best support qualitative medical education research and allow more effective methods of critical assessment of scientific literature is also needed. Advances in AI automation of administrative tasks also should be employed to reduce the burden of administrative tasks, a key contributor to burnout among academic physicians.[71] Opportunities to employ LLM applications for family medicine-specific applications (eg, focusing on the reduction of health disparities and education related to environmental and social determinants of health) should be pursued as they arise.

When choosing whether to employ cloud or local LLMs, the primary care medical education community should carefully balance the performance, cost, and security of these solutions. I call for cross-disciplinary collaboration among clinicians, educators, computer scientists, engineers, ethicists, and experts in diversity, equity, and inclusion to together build the next iteration in medical education, powered by the AI age.

FOOTNOTE

*This estimate relies on several assumptions. When LLMs are trained, each parameter is typically expressed in a 16- or 32-bit floating point number. Experience with LLMs that can be run on local (as opposed to server) hardware, such as the Llama models from Meta, suggests that quantization of the parameters down to merely 4 bits per parameter does not result in significant loss of performance. Thus, each parameter encodes about a half-byte of data, which must be used to encode information about the structure of human languages and knowledge about the world.

ACKNOWLEDGMENTS

The author thanks Dr Laurel Witt and Dr Kathryn Rampon, both at the University of Kansas Medical Center, for fruitful discussions on applying LLMs to medical education.

REFERENCES

1. Muttappallymyalil J, Mendis S, John LJ, Shanthakumari N, Sreedharan J, Shaikh RB. Evolution of technology in teaching: Blackboard and beyond in medical education. *Nepal J Epidemiol.* 2016;6(3):588-592.
2. Bostrom N. Nanoscale: Issues and Perspectives for the Nano Century, In: N C, ME M, eds. Technological revolutions: ethics and policy in the dark. Wiley; 2007:129-152.
3. Laws D. 13 Sextillion & counting: the long & winding road to the most frequently manufactured human artifact in history. *Computer History Museum.* 2018. https://computerhistory.org/blog/13-sextillion-counting-the-long-winding-road-to-the-most-frequently-manufactured-human-artifact-in-history.
4. 1947: invention of the point-contact transistor. *Computer History Museum.* 2023. https://www.computerhistory.org/siliconengine/invention-of-the-point-contact-transistor.
5. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach, 2nd ed.* Prentice Hall/Pearson; 2003.
6. Vaswani A, Shazeer N, Parmar N. Attention is all you need, 31st Annual Conference on Neural Information Processing Systems. 2017.
7. Bishop CM. *Neural Networks for Pattern Recognition.* Oxford Academic; 1995.
8. Amatriain X, Sankar A, Bing J, Bodigutla PK, Hazen TJ, Kazi M. *Transformer models: an introduction and catalog.* https://doi.org/10.48550/arXiv.2302.07730.
9. Brown T, Mann B, Ryder N. Language models are few-shot learners, 34th Annual Conference on Neural Information Processing Systems. 2020.
10. Neelakantan A, Xu T, Puri R. Text and code embeddings by contrastive pre-training. *arXiv.* 2023:2302.07730. https://doi.org/10.48550/arXiv.2201.10005.
11. Stiennon N, Ouyang L, Wu J. Learning to summarize with human feedback, 37th Annual Conference on Neural Information Processing Systems. 2023.
12. Ouyang L, Wu J, Jiang X. Training language models to follow instructions with human feedback. *arXiv.* 2023. https://doi.org/10.48550/arXiv.2203.02155.
13. Agrawal A, Gans J, Goldfarb A. ChatGPT and how AI disrupts industries. *Harvard Business Review.* 2022. https://hbr.org/2022/12/chatgpt-and-how-ai-disrupts-industries.
14. Wunker S. Disruptive innovation and ChatGPT-three lessons from the smartphone's emergence. *Forbes.* 2023. https://www.forbes.com/sites/stephenwunker/2023/02/16/disruptive-innovation-and-chatgpt--three-lessons-from-the-smartphones-emergence/?sh=7d07d7fb61aa.
15. Warzel C. Is this the week AI changed everything? *The Atlantic.* 2023. https://www.theatlantic.com/technology/archive/2023/02/google-bing-race-to-launch-ai-chatbot-powered-search-engines/673006.
16. OpenAI. GPT-4 technical report. *arXiv.* 2023:2303.08774. https://doi.org/10.48550/arXiv.2303.08774.
17. Kung TH, Cheatham M, Medenilla A. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):198.
18. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *arXiv.* 2022:2205.11916. https://doi.org/10.48550/arXiv.2205.11916.
19. Chen M, Tworek J, Jun H. Evaluating large language models trained on code. *arXiv.* 2021:2107.03374. https://doi.org/10.48550/arXiv.2107.03374.

20. Sun W, Shi Z, Gao S, Ren P, De Rijke M, Ren Z. Contrastive learning reduces hallucination in conversations. *arXiv.* 2022. https://doi.org/10.48550/arXiv.2212.10400.
21. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. *Critical Care.* 2023;27(1):120.
22. Roller S, Dinan E, Goyal N. Recipes for building an open-domain chatbot. *arXiv.* 2020:2004.13637. https://doi.org/10.48550/arXiv.2004.13637.
23. Shuster K, Poff S, Chen M, Kiela D, Weston J. Retrieval augmentation reduces hallucination in conversation. *arXiv.* 2024:2104.07567. https://doi.org/10.48550/arXiv.2104.07567.
24. Maleki N, Padmanabhan B, Dutta K. AI hallucinations: a misnomer worth clarifying. *arXiv.* 2024:2401.06796. https://doi.org/10.48550/arXiv.2401.06796.
25. Smedley BD, Stith AY, Nelson AR, Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. National Academies Press; 2003.
26. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet.* 2017;389(1):30569.
27. Freeman J. Something old, something new: the syndemic of racism and COVID-19 and its implications for medical education. *Fam Med.* 2020;52(9):623-625.
28. Sexton SM, Richardson CR, Schrager SB. Systemic racism and health disparities: a statement from editors of family medicine journals. *Ann Fam Med.* 2021;19(1):2-3.
29. Washington JC, Rodríguez JE. Racism education is needed at all levels of training. *Fam Med.* 2018;50(9):711-712.
30. Acosta D, Ackerman-Barger K. Breaking the silence: time to talk about race and racism. *Acad Med.* 2017;92(3):285-288.
31. Brooks KC. A silent curriculum. *JAMA.* 2020;323(17):690-691.
32. Ahmad NJ, Shi M. The need for anti-racism training in medical school curricula. *Acad Med.* 2017;92(8):1073.
33. Afolabi T, Borowsky HM, Cordero DM. Student-led efforts to advance anti-racist medical education. *Acad Med.* 2021;96(6):802-807.
34. Cary MP, Zink A, Wei S. Mitigating racial and ethnic bias and advancing health equity in clinical algorithms: a scoping review. *Health Aff (Millwood).* 2023;42(10):368.
35. Gervasi SS, Chen IY, Smith-Mclallen A. The potential for bias in machine learning and opportunities for health insurers to address it. *Health Aff (Millwood).* 2022;41(2):212-218.
36. Amin KS, Forman HP, Davis MA. Even with ChatGPT, race matters. *Clin Imaging.* 2024;109:110113.
37. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447-453.
38. Meade N, Poole-Dayan E, Reddy S. *An empirical survey of the effectiveness of debiasing techniques for pre-trained language models*. https://doi.org/10.48550/arXiv.2110.08527.
39. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nat Med.* 2021;27(1):136-140.
40. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial bias in pulse oximetry measurement. *N Engl J Med.* 2020;383(25):477-479.
41. Diao JA, Inker LA, Levey AS, Tighiouart H, Powe NR, Manrai AK. In search of a better equation-performance and equity in estimates of kidney function. *N Engl J Med.* 2021;384(5):396-399.
42. All of Us Research Program Investigators. The "All of Us" research program. *N Engl J Med.* 2019;381(7):668-676.
43. Mapes BM, Foster CS, Kusnoor SV. All of Us Research Program. Diversity and inclusion for the All of Us research program: a scoping review. *PLoS One.* 2020;15(7):234962.
44. Zmigrod R, Mielke SJ, Wallach H, Cotterell R. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, 57th Annual Meeting of the Association for Computational Linguistics. 2019.
45. Webster K, Wang X, Tenney I. Measuring and reducing gendered correlations in pre-trained models. *arXiv.* 2020:2010.06032. https://doi.org/10.48550/arXiv.2010.06032.
46. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. QLoRa: efficient finetuning of quantized LLMs. *arXiv.* 2023:2305.14314. https://doi.org/10.48550/arXiv.2305.14314.
47. Frantar E, Ashkboos S, Hoefler T, Alistarh D. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv.* 2022:2210.17323. https://doi.org/10.48550/arXiv.2210.17323.
48. Gerganov G. ggml. *GitHub.* 2023. https://github.com/ggerganov/ggml.
49. Open LLM leaderboard. *Hugging Face.* 2023. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
50. Gao L, Tow J, Biderman S. A framework for few-shot language model evaluation. *Zenodo.* 2021. https://zenodo.org/records/5371629.
51. Zheng L, Chiang WL, Sheng Y. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, 37th Annual Conference on Neural Information Processing Systems . 2023.
52. Kim JK, Chua M, Rickard M, Lorenzo A. ChatGPT and large language model (LLM) chatbots: the current state of acceptability and a proposal for guidelines on utilization in academic medicine. *J Pediatr Urol.* 2023;19(5):598-604.
53. Richardson C. Why ChatGPT should not be used to write academic scientific manuscripts for publication. *Ann Fam Med.* 2023:2958.
54. Stempel J. NY Times sues OpenAI, Microsoft for infringing copyrighted works. *Reuters.* 2023. https://www.reuters.com/legal/transactional/ny-times-sues-openai-microsoft-infringing-copyrighted-work-2023-12-27.
55. Beal MD, Kinnear J, Anderson CR, Martin TD, Wamboldt R, Hooper L. The effectiveness of medical simulation in teaching medical students critical care medicine: a systematic review and meta-analysis. *Simul Healthc.* 2017;12(2):104-116.
56. Maloney S, Haines T. Issues of cost-benefit and cost-effectiveness for simulation in health professions education. *Advances in Simulation.* 2016;1:13-13.
57. Peng S, Kalliamvakou E, Cihon P, Demirer M. The impact of AI on developer productivity: evidence from GitHub copilot. *arXiv.* 2023:2302.06590. https://doi.org/10.48550/arXiv.2302.06590.
58. Bowen JL, Irby DM. Assessing quality and costs of education in the ambulatory setting: a review of the literature. *Acad Med.* 2002;77(7):621-680.

59. Irby DM. Teaching and learning in ambulatory care settings: a thematic review of the literature. *Acad Med.* 1995;70(10):898-931.
60. Ridder JVD, Mcgaghie WC, Stokking KM, Cate OT. Variables that affect the process and outcome of feedback, relevant for medical training: a meta-review. *Med Educ.* 2015;49(7):658-673.
61. Warm E, Kelleher M, Kinnear B, Sall D. Feedback on feedback as a faculty development tool. *J Grad Med Educ.* 2018;10(3):354-355.
62. Danilovich N, Kitto S, Price DW, Campbell C, Hodgson A, Hendry P. Implementing competency-based medical education in family medicine: a narrative review of current trends in assessment. *Fam Med.* 2021;53(1):9-22.
63. Pelgrim E, Kramer A, Mokkink H, Van Der Vleuten C. Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. *BMC Medical Education.* 2012;12:97.
64. Bradley EA, Winchester D, Alfonso CE. Physician wellness in academic cardiovascular medicine: a scientific statement from the American Heart Association. *Circulation.* 2022;146(16):229-241.
65. Dillon EC, Tai-Seale M, Meehan A. Frontline perspectives on physician burnout and strategies to improve well-being: interviews with physicians and health system leaders. *J Gen Intern Med.* 2020;35(1):261-267.
66. Jordan J, Clarke SO, Coates WC. A practical guide for conducting qualitative research in medical education: Part 1-How to interview. *AEM Educ Train.* 2021;5(3):10646.
67. Pope C, Ziebland S, Mays N. Analysing qualitative data. *BMJ.* 2000;320(7227):114-116.
68. Xiao Z, Yuan X, Liao QV, Abdelghani R, Oudeyer PY. Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. *arXiv.* 2023:2304.10548. https://doi.org/10.48550/arXiv.2304.10548.
69. Koehl D, Vangsness L. Measuring latent trust patterns in large language models in the context of human-AI teaming, 67th Human Factors and Ergonomics Society Annual Meeting . 2023.
70. Hake J, Crowley M, Coy A. Quality, accuracy, and bias in ChatGPT-based summarization of medical abstracts. *Ann Fam Med.* 2024;22(2):113-120.
71. Rao SK, Kimball AB, Lehrhoff SR. The impact of administrative burden on academic physicians: results of a hospital-wide physician survey. *Acad Med.* 2017;92(2):237-243.