Large Language Models and Medical Education: Preparing for a Rapid Transformation in How Trainees Will Learn to Be Doctors

Akshay Ravi[1], Aaron Neinstein[1,2], and Sara G. Murray[1,3]

[1]Department of Medicine, [2]Center for Digital Health Innovation and [3]Health Informatics, University of California, San Francisco, San Francisco, California

ORCID ID: 0000-0002-4346-0555 (A.R.)

ABSTRACT

Artificial intelligence has the potential to revolutionize health care but has yet to be widely implemented. In part, this may be because, to date, we have focused on easily predicted rather than easily actionable problems. Large language models (LLMs) represent a paradigm shift in our approach to artificial intelligence because they are easily accessible and already being tested by frontline clinicians, who are rapidly identifying possible use cases. LLMs in health care have the potential to reduce clerical work, bridge gaps in patient education, and more. As we enter this era of healthcare delivery, LLMs will present both opportunities and challenges in medical education. Future models should be developed to support trainees to develop skills in clinical reasoning, encourage evidence-based medicine, and offer case-based training opportunities. LLMs may also change what we continue teaching trainees with regard to clinical documentation. Finally, trainees can help us train and develop the LLMs of the future as we consider the best ways to incorporate LLMs into medical education. Ready or not, LLMs will soon be integrated into various aspects of clinical practice, and we must work closely with students and educators to make sure these models are also built with trainees in mind to responsibly chaperone medical education into the next era.

Keywords:
artificial intelligence; large language models; medical education

(Received in original form March 19, 2023; accepted in final form June 1, 2023)

ATS Scholar Vol 4, Iss 3, pp 282–292, 2023

DOI: 10.34197/ats-scholar.2023-0036PS

CHALLENGES IMPLEMENTING ARTIFICIAL INTELLIGENCE IN HEALTH CARE

Artificial intelligence (AI) has already started to transform the world around us, but, until now, health care has lagged significantly behind other industries (1). Although an extraordinary number of machine learning models have been built in health care, only a small proportion of these models have actually been implemented (2–4). Part of the reason for the disparity between model generation and implementation is that current efforts in AI are aimed at the wrong use cases and may not always be designed ethically.

In many cases, the problems are chosen on the basis of what can be most easily predicted rather than where we need AI to drive improvement. When we focus on the easiest thing to predict, we may be able to build models with reasonable performance characteristics, but often we cannot translate those predictions into meaningful actions (2). An algorithm predicting hospital readmissions may perform well but may have limited clinical utility if it cannot identify preventable readmissions or highlight modifiable risk factors to reduce the risk of readmission (5). In choosing these problems, we often neglect use cases such as customer support or back-office process optimization, which are in fact the most common areas for AI adoption across other industries (6).

Even predictive models built by companies such as major electronic health record (EHR) vendors with the intent of operationalizing them have failed to deliver robust and ethical solutions. We have found that some industry-developed models perform poorly, have limited external validity, or risked serious bias against vulnerable patient populations (7–10).

A TRANSFORMATIVE PARADIGM SHIFT

Large language models (LLMs) are semisupervised, generative transformer language models that are trained on massive amounts of text data and effectively contextualize the sequential nature of words in a sentence to predict the most plausible response. Essentially, they are able to interpret user prompts and provide a conversational response, such as the case with the recently popularized ChatGPT, developed by OpenAI, as well as other tasks, including summarization and translation (11). Although we are just beginning to scratch the surface of potential use cases in health care, many are speculating that these tools have transformative potential (3, 12).

Because these tools are widely available to the public, they allow a paradigm shift in how we think about identifying use cases for AI in medicine. Testing use cases with simple prompt engineering does not require the same data science resources that are needed to build a model from scratch. Doctors are already testing the potential for LLMs in medicine by asking tools such as ChatGPT to draft insurance appeals or preauthorization letters, summarize patient instructions at varying reading levels, generate differential diagnoses, and more (13–16). Many of these use cases, if deployed in a trustworthy and secure way, may serve to improve daily work.

POSSIBILITIES AND PITFALLS

LLMs, if implemented well and with careful consideration of their limitations,

Correspondence and requests for reprints should be addressed to Akshay Ravi, M.D., Department of Medicine, University of California, San Francisco, 521 Parnassus Avenue, Box 0131, San Francisco, CA 94143. E-mail: akshay.ravi@ucsf.edu.

Table 1. Sample use cases for large language models in health care

Use Case	Possibilities
Task automation	• Drafting prior authorization requests, insurance appeals letters, work letters, disability paperwork, medical leave requests, and more • Triaging patient messages (in-basket) and drafting a skeleton response for provider review
Patient education	• Translating medical jargon into clinic visit instructions or discharge instructions at a patient-friendly reading level • Drafting lifestyle counseling recommendations such as diet, exercise, basic physical therapy etc. • Creating medication tables with clear instructions for use and side effects
Clinical documentation	• Using prior clinic visit notes, diagnoses, laboratory test results, and imaging results to prepare the outlines of a note • Summarizing long hospital course to prepare an outline of a discharge summary • Pulling out key information from a note to identify billing level
Diagnostic assistance	• Offering a basic differential diagnosis based on a problem representation or the available clinical data in the EHR • Identifying case reports of patients with similar problems or rare diseases
Literature review	• Summarizing key studies and their findings in a particular field • Identifying and summarizing the articles relevant to a new research project • Copyediting academic manuscripts to help bridge equity gaps for trainees and providers for whom English is a second language

Definition of abbreviation: EHR = electronic health record.

could enhance patient care, dramatically reduce clerical work, and improve both patient and provider satisfaction. Some of the potential use cases for LLMs in health care are presented in Table 1 and summarized below.

Reducing clerical work could dramatically improve daily physician work. By targeting task automation or clinical documentation, these models could enable physicians to spend less time behind a screen and more time with their patients. With rising physician burnout related to growing documentation burdens and administrative tasks, LLMs represent a great opportunity to lighten this load (17–19). If trained on medical record data, future iterations of these models could provide drafts of prior authorizations or family medical leave letters that are tailored to the patient. Similarly, these future LLMs could prepare an outline of a clinical note before any patient encounter, including relevant past visits, laboratory test results, imaging, and more, to effectively "preround" for the doctor. Eventually, such models could be implemented in concert with ambient, automatic audio recordings of clinic visit encounters to serve as an "electronic

scribe" and prepare a structured note by the end of the visit for a clinician to review before signing. Although previous iterations of digital scribes have had issues with accuracy and appropriate formatting for clinical notes, future language models that are trained using reinforcement learning with human feedback can optimize these models to understand and translate human conversation into clinically useful notes (20). More importantly, they may mitigate physician responsibilities for billing by interpreting and coding notes on the basis of what has already been documented. These possibilities could help offload a substantial amount of clerical work for physicians, allowing them to get back to what they care about: delivering high-quality patient care.

LLMs could also play a key role in bridging gaps in patient education. Patient communication and education have been recognized as vital to providing safe patient care; yet, there are still barriers to effective patient communication, especially for patients with limited English proficiency (21–24). LLMs could potentially serve as translators, taking in complex medical jargon and simplifying it for patients at varying reading levels and in various languages. These could apply to clinic visit summaries, discharge instructions, lifestyle counseling, medication counseling, and more to improve health equity in communication for vulnerable patients. Once trained on medical domain–specific data or EHR data, these models could even provide a first draft of patient instructions based on a prompt as simple as "blood pressure monitoring instructions" and provide summaries that differ on the basis of the reading level of the patient or in the patient's native language. Such interventions not only would improve provider efficiency but also could dramatically improve patients' understanding of their health.

However, for all of these use cases, we must be cautious of the limitations of LLMs. The U.S. Department of Health and Human Services has released guidelines for six key principles of trustworthy AI that should be considered before implementing any AI tool in health care and can be a useful frame for evaluating LLMs (25):

1. *Robust/reliable*: LLMs must be accurate and may need additional training on healthcare-specific data to allow them to minimize incorrect information conveyed. They will also need regular retraining to incorporate new studies, guidelines, and recommendations. Considering these limitations, even OpenAI has disallowed the use of its models to provide diagnostic or therapeutic services as of March 23, 2023 (26).
2. *Fair/impartial*: LLMs must be trained thoughtfully to prevent the incorporation of explicit or implicit biases found in their training data, which could perpetuate socioeconomic disparities in health care (27, 28). Efforts to mitigate these biases are ongoing, including the development of libraries to measure and describe model bias (29).
3. *Transparent/explainable*: The output of LLMs must be explainable or justifiable such that clinicians using them can reasonably understand the response provided. Currently, LLMs cannot refer to their source materials to justify a response or provide a level of certainty with any given answer and will need these features in future iterations.
4. *Responsible/accountable*: LLMs will need strict protocols of review and oversight by clinicians to ensure that safe and effective patient care is provided.
5. *Safe/secure*: LLMs trained on patient data must protect that data. As future iterations of these models are built incorporating EHR data, safeguards will be needed to ensure the safety of these data and prevent data leak (30).

6. *Privacy/consent*: LLMs must be built ethically with patient preference in mind. Trust is key in the doctor–patient relationship, and the use of LLMs to draft responses to patient messages could erode this trust, especially in sensitive spaces such as mental health counseling (31). Future LLMs that are built on EHR data will need explicit consent from patients.

Finally, it is unclear how LLMs should be credited, as exemplified by current controversies in the use of AI-generated text in academic writing (32). In this and other contexts, a middle path may be appropriate in which the use of LLMs is acknowledged but does not rise to the level of accountability needed for true authorship credit (33). Examples of how these principles of trustworthy AI can be applied to LLMs are shown in Table 2. For these reasons, it is critical that LLMs in their current form be used only to streamline physician work rather than replace it.

TRAINEES AND LLMS

As we enter a new era of healthcare delivery in which LLMs will be readily available to our trainees, we anticipate there will be both great opportunities and challenges in how these models will impact the next generation of physicians and the ways in which we teach them.

Coaching for Synthesizing Patient Data into an Assessment and Plan

One of the most challenging leaps for trainees is the transition from a preclerkship to a clerkship style of education (34). Students must learn to consolidate and synthesize vast amounts of medical knowledge into relevant clinical knowledge for the patient in front of them. They must process real-world data to produce problem representations, differential diagnoses, and plausible plans for a patient. Traditionally, early learners work closely with their supervising residents or attending physician to hone these skills. However, these interactions may sometimes be performative and high stakes for the learner, which may not suit all trainees. LLMs could serve as a low-stakes alternative to support trainees in self-directed learning during this process. Imagine a future LLM trained on clinical note text, discrete data, PubMed, and other medical references that is able to review draft clinical notes from trainees and provide additional recommendations to their differential diagnoses or plans and explain why with helpful references (35). Although this certainly cannot replace clinical expertise, it can be one more helpful nudge in the right direction. However, it will be important to consider best practices for how learners interact with these types of models. An AI model that reviews a student's work and provides feedback can be helpful in consolidating knowledge, but a model that does the thinking for them could be harmful instead. Furthermore, ensuring the validity of LLM responses for various use cases will be critical before learners engage with potentially erroneous output.

Facilitating the Effective Application of Evidence-based Practice

In the era of evidence-based medicine, trainees are taught to actively incorporate the practice of seeking and appraising medical literature in the care of patients (36, 37). This includes defining the problem, recognizing the information that might solve this problem, searching the literature for relevant studies, interpreting the results of these studies, and applying them to the care of the patient at hand (37). Although the process of evaluating a study and, critically, deciding whether it could apply to your patient is still a vital

Table 2. Principles of trustworthy artificial intelligence applied to large language models

Principles of Trustworthy AI	Specific Examples for LLMs
Robust/reliable	• Current LLMs are at risk of "hallucination," or providing plausible-sounding but incorrect information. For example, some current LLMs will cite studies that sound realistic but do not exist. This is an evolving problem, and LLMs themselves may be able to help identify hallucination (44). • Some LLMs may have knowledge cutoffs, or date limits on the most recent data that have been used to train the model. Some periodic retraining of the model would be necessary to incorporate new studies, guidelines, and recommendations as they arise.
Fair/impartial	• LLMs can incorporate biases found in their training data, which could inadvertently perpetuate harmful racial, sex, and other biases (27, 28). For example, this could lead to bias in generating a work letter for male versus female-identifying patients or in drafting a response to a patient portal message about pain for White versus Black patients.
Transparent/explainable	• Current publicly available LLMs are not able to provide much explainability in their responses in the form of either references to the source materials that they are using to formulate a response or providing an assessment of certainty in the accuracy of a response. This limits a user's ability to accurately interpret a model's responses.
Responsible/accountable	• Because these tools are still in their infancy, there will need to be strict supervision and oversight from the physicians who use LLMs to make sure that information conveyed by these models is not inaccurate or incomplete. For example, although LLMs may be able to draft patient portal message responses, a physician still needs to review and read this message before sending it.
Safe/secure	• Current LLMs have not been built using any specific patient data, but already they have had issues with the leak of conversations between users (30), which is why both physicians and healthcare organizations must be cautious with their use. Physicians should not use protected health information through unsecured online LLMs, and healthcare organizations should create systems for secure computing and business associate agreements to ensure the safety of these data if partnering with organizations that build LLMs.
Privacy/consent	• Publicly available LLMs have been trained on public internet data, but future LLMs built on electronic health record data will need explicit approval and consent from patients regarding the use of their data. • Patients may not be comfortable with messages or documentation from their providers being drafted wholly or in part by a model. This can have serious implications for patient trust, especially if used for sensitive topics (31). Such technology should allow patients to opt in or opt out.

Definition of abbreviations: AI = artificial intelligence; LLM = large language model.
Adapted from Reference 25.

part of training, the process of searching for the evidence base is not. Although current LLMs may not be capable of citing a source within their training data, future LLMs could be trained for this purpose on a "reliable" corpus of articles like PubMed, such as BioBert or BioMedLM. These LLMs could efficiently highlight the high-quality studies most relevant to the problem, allowing trainees to focus on the appraisal step (38, 39). Furthermore, trainees could use these future LLMs to help familiarize themselves with the evidence base behind well-known practice patterns, such as in the management of patients with heart failure with a reduced ejection fraction.

Generating Synthetic Training Cases for Interactive Learning

Case-based training has been a staple of medical education. This entails authored cases with focused questions aimed at stimulating clinical reasoning and the gradual release of the relevant physical examination, laboratory tests, and imaging to help the trainee make a diagnosis and treatment plan. If trained on EHR data, an LLM could conceivably generate synthetic patient cases in a similar way and potentially even interact with trainees to reveal bits and pieces of information as the case progresses. Each case could be completely unique and challenge learners to manage diagnostic uncertainty in a safe, simulated environment. Future iterations could even be used to evaluate trainees as a novel form of examination.

Of course, there will be significant limitations of LLMs as they relate to medical education. There is justifiable concern about the accuracy of these models. Although recent studies have shown that ChatGPT and other LLMs trained on medical domain–specific data are capable of passing the United States Medical Licensing Examination, these models only pass with 60–68% accuracy (40, 41). This is an incredible performance for a language model but a paltry performance if we are to consider the tool to be an educator on the topic. Moreover, such models could present plausible-sounding answers to trainees who may not be able to discern reality from fiction. As they currently exist, it may be hard to tease out these errors because tools such as ChatGPT cannot currently justify or provide references for their response.

It is also important to consider how the implementation of these theoretical LLMs into clinical workflows may change what is important to continue teaching our trainees. For example, as ambient "scribe" LLMs become more prevalent, the basic structure of a history and physical examination or subjective, objective, assessment, and plan note may become less important to formally teach than the way to communicate an assessment and plan with the patient. Instead, it may be more valuable to teach trainees how to write appropriate prompts, validate responses, and detect potential errors. LLMs will also challenge educators to reframe their approach to clinical teaching. For example, in a world in which trainees could easily access a differential diagnosis from an LLM, clinical teaching may focus instead on the conceptual frameworks around that differential or the softer skills of clinical reasoning. As these models alleviate our nonclinical burdens, we may be able to spend more time with trainees and patients to model these skills. As our trainees may have greater familiarity and comfort with technology and may not have the workflows of today ingrained in them, they will be vital partners in envisioning a future in health care that can use LLMs to transform clinical care.

Furthermore, if trainees are involved in the training and supervision of these models, they can help develop these models to be effective teachers for future trainees.

Finally, we must consider the ethical implications of using language models. There has been widespread popular concern about the use of LLMs to cheat on examinations or other forms of evaluation. Similarly, there may be concerns about the use of LLMs by students on the wards or in research as a form of academic dishonesty. Learners who use LLMs to supplement their differential diagnosis or improve their plan may be using this tool appropriately, but those who use LLMs to generate an entire clinical note or academic article that they claim as their own may not be. The "line of appropriateness" may also shift as trainees progress from undergraduate medical education to graduate medical education. This problem is further exacerbated by the fact that it can be hard for humans to distinguish between AI-generated text and natural text (42). Future LLMs may need to be built in conjunction with AI classifiers to help educators understand how and to what extent trainees are using LLMs (42, 43). However, learners will still likely use tools such as LLMs, regardless of whether they are formally incorporated into their curricula or clinical experiences. Rather than letting medical trainees use ChatGPT or whichever LLM they choose, it would be more ideal if we provided them with sanctioned LLMs that are trained on the most appropriate corpus, implemented at the right point in the workflow, and provide the context and explanations needed to enhance learning.

CONCLUSIONS

Whether we are ready or not, LLMs such as ChatGPT, Med-PaLM, and others will soon be integrated into various aspects of clinical practice. We anticipate that they will transform patient and provider experience and medical education for our trainees. Because academic medical centers are at the forefront of clinical innovation and medical education, they must engage with stakeholders to thoughtfully design and implement these tools. We must develop key partnerships with technical leaders within our organizations and with external companies with expertise in LLMs. We must think about the privacy, safety, and security of training these models, especially if patient data are used. We must work with frontline clinicians to understand their successes and failures with LLMs and center usability in our implementations. We must work with educators and students to make sure these models are also built with trainees in mind to ethically and responsibly chaperone medical education into the next era.

Author disclosures are available with the text of this article at www.atsjournals.org.

REFERENCES

1. Brynjolfsson E, McAfee A. The business of artificial intelligence. *Harv Bus Rev* 2017 [accessed 2023 Feb 17]. Available from: https://hbr.org/2017/07/the-business-of-artificial-intelligence.

2. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. *BMJ Innov* 2020;6:45–47.

3. Zhang A, Xing L, Zou J, Wu JC. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat Biomed Eng* 2022;6:1330–1345.

4. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.

5. Liu W, Stansbury C, Singh K, Ryan AM, Sukul D, Mahmoudi E, *et al.* Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLoS One* 2020;15: e0221606.

6. Chui M, Hall B, Mayhew H, Singla A, Sukharevsky A. The state of AI in 2022—and a half decade in review. McKinsey and Co.; 2022 [accessed 2023 Feb 17]. Available from: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review.

7. Soleimani H, Murray SG. Double dice roll outperforms a built-in model for predicting remaining length of stay: lessons learned from a prospective evaluation. AMIA Annual Symposium. November 14, 2020, Virtual.

8. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, *et al.* External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med* 2021;181:1065–1070.

9. Murray SG, Wachter RM, Cucina RJ. Discrimination by artificial intelligence in a commercial electronic health record—a case study. *Health Aff Forefr* [accessed 2023 Mar 7]. Available from: https://www.healthaffairs.org/content/forefront/discrimination-artificial-intelligence-commercial-electronic-health-record-case-study.

10. Oikonomidi T, Norman G, McGarrigle L, Stokes J, van der Veer SN, Dowding D. Predictive model-based interventions to reduce outpatient no-shows: a rapid systematic review. *J Am Med Inform Assoc* 2023;30:559–569.

11. OpenAI. ChatGPT: optimizing language models for dialogue. 2022 [accessed 2023 Feb 17]. Available from: https://openai.com/blog/chatgpt/.

12. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, *et al.* Language models are few-shot learners. arXiv; 2020 [accessed 2023 Feb 17]. Available from: https://arxiv.org/abs/2005.14165.

13. Clifford Stermer, MD on TikTok. TikTok [accessed 2023 Feb 19]. Available from: https://www.tiktok.com/@tiktokrheumdok/video/7176660771806383403.

14. Gabrielson AT, Odisho AY, Canes D. Harnessing generative artificial intelligence to improve efficiency among urologists: welcome ChatGPT. *J Urol* 2023;209:827–829.

15. Shen Y, Heacock L, Elias J, Hentel KD, Reig B, Shih G, *et al.* ChatGPT and other large language models are double-edged swords. *Radiology* 2023;307:e230163.

16. Glass Health. The first digital notebook designed for doctors [accessed 2023 Mar 14]. Available from: https://glass.health/.

17. Shanafelt TD, Boone S, Tan L, Dyrbye LN, Sotile W, Satele D, *et al.* Burnout and satisfaction with work-life balance among US physicians relative to the general US population. *Arch Intern Med* 2012; 172:1377–1385.

18. West CP, Dyrbye LN, Shanafelt TD. Physician burnout: contributors, consequences and solutions. *J Intern Med* 2018;283:516–529.

19. Johnson KB, Neuss MJ, Detmer DE. Electronic health records and clinician burnout: a story of three eras. *J Am Med Inform Assoc* 2021;28:967–973.

20. Li B, Crampton N, Yeates T, Xia Y, Tian X, Truong K. Automating clinical documentation with digital scribes: understanding the impact on physicians. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. May 6, 2021, Yokohama, Japan. pp. 1–12.

21. Ryan R, Santesso N, Lowe D, Hill S, Grimshaw J, Prictor M, *et al.* Interventions to improve safe and effective medicines use by consumers: an overview of systematic reviews. *Cochrane Database Syst Rev* 2014;2014:CD007768.

22. Ho T, Campos BS, Tarn DM. Post-visit patient understanding about newly prescribed medications. *J Gen Intern Med* 2021;36:3307–3310.

23. Schoen C, Osborn R, Huynh PT, Doty M, Zapert K, Peugh J, *et al.* Taking the pulse of health care systems: experiences of patients with health problems in six countries. *Health Aff (Millwood)* 2005;24:W5-509–W5-525.

24. Mir TH, Osayande A, Kone K, Bridges K, Day P. Assessing the quality of the after-visit summary (AVS) in a primary-care clinic. *J Am Board Fam Med* 2019;32:65–68.

25. U.S. Department of Health and Human Services. Artificial intelligence (AI) at HHS. 2021 [accessed 2023 May 7]. Available from: https://www.hhs.gov/about/agencies/asa/ocio/ai/index.html.

26. OpenAI. Usage policies [accessed 2023 Apr 26]. Available from: https://openai.com/policies/usage-policies.

27. Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories. In: Proceedings of the Third Workshop on Narrative Understanding. Association for Computational Linguistics; 2021 Virtual, pp. 48–55 [accessed 2023 Apr 21]. Available from: https://aclanthology.org/2021.nuse-1.5.

28. Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. *Nat Mach Intell* 2021;3:461–463.

29. Hugging Face. Evaluating language model bias with evaluate [accessed 2023 Apr 21]. Available from: https://huggingface.co/blog/evaluating-llm-bias.

30. OpenAI. March 20 ChatGPT outage: here's what happened [accessed 2023 Apr 25]. Available from: https://openai.com/blog/march-20-chatgpt-outage.

31. Ingram D. AI Chat used by mental health tech company in experiment on real users. *NBC News.* 2023 [accessed 2023 Apr 25]. Available from: https://www.nbcnews.com/tech/internet/chatgpt-ai-experiment-mental-health-tech-app-koko-rcna65110.

32. International Conference on Machine Learning (ICML). Clarification on large language model policy LLM. 2023 [accessed 2023 Apr 27]. Available from: https://icml.cc/Conferences/2023/llm-policy.

33. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. *Nature* 2023;613:612.

34. Poncelet A, O'Brien B. Preparing medical students for clerkships: a descriptive analysis of transition courses. *Acad Med* 2008;83:444–451.

35. Yang X, Chen A, PourNejatian N, Shin HC, Smith KE, Parisien C, *et al.* A large language model for electronic health records. *NPJ Digit Med* 2022;5:194.

36. Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. *Lancet* 2017;390:415–423.

37. Guyatt G, Cairns J, Churchill D, Cook D, Haynes B, Hirsh J, *et al.*; Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA* 1992;268:2420–2425.

38. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020;36:1234–1240.

39. Venigalla A, Frankle J, Carbin M. BioMedLM: a domain-specific large language model for biomedical text. Mosaic [accessed 2023 Mar 10]. Available from: https://www.mosaicml.com/blog/introducing-pubmed-gpt.

40. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, *et al.* Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS Digit Health* 2023;2:e0000198.

41. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, *et al.* Large language models encode clinical knowledge. arXiv; 2022 [accessed 2023 Feb 19]. Available from: https://arxiv.org/abs/2212.13138.

42. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, *et al.* Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv; 2022 [accessed 2023 Apr 26]. Available from: https://www.biorxiv.org/content/10.1101/2022.12.23.521610v1.

43. OpenAI. New AI classifier for indicating AI-written text [accessed 2023 Apr 26]. Available from: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

44. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med* 2023;388:1233–1239.