Potential risks of GenAI on medical education

Jacob Hough,[1] Nicholas Culley,[1] Chase Erganian,[1] Fares Alahdab [2]

10.1136/bmjebm-2025-114339

[1]School of Medicine, University of Missouri, Columbia, Missouri, USA
[2]Department of Biomedical Informatics, Biostatistics, and Medical Epidemiology, University of Missouri School of Medicine, Columbia, Missouri, USA

Correspondence to: **Dr Fares Alahdab**; faris.ahdab@gmail.com

To cite: Hough J, Culley N, Erganian C, *et al*. *BMJ Evidence-Based Medicine* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjebm-2025-114339

Generative artificial intelligence (AI) moved from an experimental novelty into mainstream use as an everyday tool in under 3 years. Large language models (LLMs) now pass portions of knowledge examinations and draft clinically relevant text that gives the impression of being comparable to human output.[1–3] These tools are likely already embedded in student study habits, faculty workflows and clinical communication. Recent survey data show widespread use among learners for content review, homework assignments, draft writing and feedback, with largely insufficient institutional policies and guidance.[4–7] Much literature focuses on the benefits and promises of AI models in healthcare. In this piece, we aim to highlight potential risks and provide guidance on mitigating them, with a focus on medical education.

Risks of over-reliance on AI

Due to the widespread use across a vast array of tasks, along with the burgeoning potential, the risks stemming from over-reliance on these tools are growing proportionally. AI tools may accordingly amplify known human factors problems in health professions education. Medical school and training programmes need to consider risks such as automation bias, cognitive off-loading and outsourcing of reasoning, de-skilling (with the greatest harm to novices), bias and inequity, hallucinated content and sources, and privacy, security and data governance.

Automation bias is when, after extended use and over-reliance, trainees accept incorrect recommendations from AI systems. Generic warnings against this do not prevent it, since the root cause is over-trust. Medical training should include practice in rejecting poor AI advice and in explaining why it is unsafe to follow.[8 9] Relatedly, *cognitive offloading and outsourcing of reasoning capacities* are a risk that creeps in silently and undetected at first.[10] Fluent AI model outputs may tempt users to skip independent information retrieval, appraisal and synthesis. This may also harm critical thinking capacities in learners. When this goes on for a long time, the consequence of *de-skilling* emerges, where skills deteriorate rather than remain sharp.[11–13] What happens if the servers or AI services go down? The impact of this is particularly ominous for learners who are working on developing the skill in the first place, as they are denied the opportunity to do so in the process (table 1).

The potential negative impact of AI in medical education also extends to emphasising existing biases in the data on which the models were trained.[14] AI models have reproduced *racial and other demographic biases*,[15 16] and we need to emphasise this risk and take it into account in both coursework and assessment. *Hallucinating confident falsehoods and sources* remains a frequent failure mode for AI models.[17 18] This requires careful and critical assessment of the model outputs by the learners, along with learning how to search for, retrieve, understand and assess relevant evidence-based information to measure against the models' responses. Finally, *risks to personal privacy, security and data ownership* exist with the use of commercially available AI models, particularly ones that are opaque about their data retention policies or that offer few options for users' data control. Given the sensitive nature of the healthcare domain, such risks are particularly high, including for learners as well.

Charting a safe way forward

The goal of using AI to augment education, rather than letting it erode independent reasoning, is a worthy pursuit. As AI is disrupting traditional learning and evaluation methods, adjustments to medical school and training curricula are necessary. For educational assessments, first, it may be wise to assume learners' access to and use of AI and to grade the process, rather than only the product. This can be done by asking the students to 'show their work', provide a paper trail and even submit the LLM prompts they used along with written rationales for accepting or rejecting the AI's output. Second, for critical skills assessments, we need to design AI-free assessments using supervised stations or in-person examinations. This may be feasible, and especially important, for bedside communication, physical examination, teamwork and professional judgement. Third, it may be important to evaluate AI itself as a competency. Data literacy and teaching AI design, development and evaluation are more important now than ever, and this knowledge is no longer a luxury for medical learners and trainees. Medical trainees may not need to be fully emerged into the technical data engineering details and training pipelines for AI models, but they should understand that process in principle and grasp the concepts underpinning its strengths and weaknesses. They should also understand where and how such AI tools can be embedded in clinical workflows and care pathways, and how to evaluate their intended performance and potential biases over time. Just as in evidence-based medicine, skills and education that enable learners to build expertise in finding and appraising clinical evidence can also be appropriately expanded and applied to research on AI in healthcare. This rapidly growing body of knowledge underscores the need

Table 1 Definitions and suggested mitigation strategies for potential risks of genAI on medical education

Risk	Detail	Example	Mitigation strategy
Automation bias	Tendency to over-trust automated recommendations even when they are wrong.	During a case-based quiz, students accept an LLM-suggested differential and management plan without reconciling contradictions in the vignette.	Build 'trust-calibration' labs that mix correct and flawed AI outputs; require a documented disconfirming evidence search and/or oral defence; grade process artefacts (prompts, rationale, verification).
Cognitive off-loading and outsourcing reasoning	Shifting information retrieval, appraisal and synthesis to AI so that learners generate fewer internal explanations and weaker memory traces.	First-year students use AI to write pathophysiology explanations; in viva voce, they cannot reproduce the reasoning chain or sources.	Use AI-permitted assignments that require retrieval practice and concept maps authored by the student, plus brief viva checks; alternate AI-enabled with AI-free tasks to sustain effortful encoding.
De-skilling	Loss of procedural or cognitive skill from premature automation of routine tasks; experts may compensate, novices cannot.	Students rely on AI note templates and stop writing full differentials.	Set minimum competence thresholds before AI use is allowed for a skill; maintain AI-free OSCE stations and manual calculations; require periodic 'skill re-demonstration' with faculty feedback.
Racial and equity biases	Systematic differences in model outputs by protected characteristics or propagation of race-based biases.	An AI-generated case suggests different analgesia choices for otherwise identical black vs white patients; biased descriptors appear in autodrafted notes.	Bake bias audits into coursework: counterfactual testing across race, gender and language; track subgroup metrics.
Hallucinations	Fluent but false statements or references produced with high confidence.	A literature review assignment includes an AI-invented randomised trial and non-existent citations.	Mandate primary-source verification and reference cross-checks; grade a 'verification step' explicitly; use reporting checklists for chatbot-assisted work and require disclosure consistent with scholarly policies.
Privacy, security and data governance	Risks include exposing PHI or student records, prompt injection, data exfiltration and uncontrolled retention of sensitive content.	A student pastes a 'de-identified' patient note into a public chatbot that still contains dates and rare disease details; content is retained outside school control.	Enforce a no-PHI/no-personally identifiable information policy for consumer tools; use institutionally hosted models with logging and retention controls.

AI, artificial intelligence; LLM, large language model; OSCE, objective structured clinical examination; PHI, protected health information.

to be able to read, understand, appraise and act on this evidence in an informed manner. This should include education on uncertainty communication, bias checks, AI limitations and nuances of clinical workflow AI deployment. Enhanced critical thinking teaching is especially needed, which can be achieved by building cases where the AI outputs are a mix of correct and intentionally flawed responses (ie, 'trust calibration' laboratory studies). Learners would then accept, amend or reject, and justify their decision with primary evidence-based sources. Similarly, building modules that showcase the AI risks to equities and fairness can be done by matching cases that differ only in protected characteristics. Students then probe for output differences and design mitigations, while citing bias literature to justify their choices.

Generative AI has documented and well-researched benefits, but it is not without pitfalls, particularly to medical education and novice learners. These tools can fabricate sources, encode bias, lead to over-reliance and have negatively disruptive effects on the educational journey. Medical programmes must be vigilant about these risks and adjust their curricula and training programmes to stay ahead of them and mitigate their likelihood. They should also use competency-based frameworks for the continuous evaluation of these potential risks and adapting the curricula, content and assessments to maintain high fidelity of the educational process. Finally, regulatory bodies, professional societies and educational associations, such as the World Medical Association, World Federation for Medical Education, International Association for Health Professions Education and their counterparts in different regions and countries of the world, including the Association for American Medical Colleges and Association for Medical Education in Europe, should maintain updated evidence-based guidance on the impact of AI on medical education. The field also requires the active involvement and support of these important entities in the continuous evaluation and research effort to study the benefits and risks of AI on medical education.

Contributors FA: conceptualisation and manuscript writing. JH, NC, CE: manuscript writing.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD
Fares Alahdab https://orcid.org/0000-0001-5481-696X

References

1 **Shaikh Y**, Jeelani-Shaikh ZA, Jeelani MM, *et al*. Collaborative intelligence in AI: Evaluating the performance of a council of AIs on the USMLE. *PLOS Digit Health* 2025;4:e0000787.
2 **Preiksaitis C**, Rose C. Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review. *JMIR Med Educ* 2023;9:e48785.
3 **Van Veen D**, Van Uden C, Blankemeier L, *et al*. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med* 2024;30:1134–42.

4 **McCoy L**, Ganesan N, Rajagopalan V, *et al*. A Training Needs Analysis for AI and Generative AI in Medical Education: Perspectives of Faculty and Students. *J Med Educ Curric Dev* 2025;12:23821205251339226.
5 **Sakelaris PG**, Novotny KV, Borvick MS, *et al*. Evaluating the Use of Artificial Intelligence as a Study Tool for Preclinical Medical School Exams. *J Med Educ Curric Dev* 2025;12:23821205251320150.
6 **Ichikawa T**, Olsen E, Vinod A, *et al*. Generative Artificial Intelligence in Medical Education—Policies and Training at US Osteopathic Medical Schools: Descriptive Cross-Sectional Survey. *JMIR Med Educ* 2025;11:e58766.
7 **Banerjee M**, Chiew D, Patel KT, *et al*. The impact of artificial intelligence on clinical education: perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. *BMC Med Educ* 2021;21:429.
8 **Jabbour S**, Fouhey D, Shepard S, *et al*. Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study. *JAMA* 2023;330:2275–84.
9 **Khera R**, Simon MA, Bias RJSA, *et al*. Risk of Harm From AI-Driven Clinical Decision Support. *JAMA* 2023;330:2255–7.
10 **Jiang T**, Wu J, Leung SCH. The cognitive impacts of large language model interactions on problem solving and decision making using EEG analysis. *Front Comput Neurosci* 2025;19:1556483.
11 **Choudhury A**, Chaudhry Z. Large Language Models and User Trust: Consequence of Self-Referential Learning Loop and the Deskilling of Health Care Professionals. *J Med Internet Res* 2024;26:e56764.
12 **Wadhwa V**, Alagappan M, Gonzalez A, *et al*. Physician sentiment toward artificial intelligence (AI) in colonoscopic practice: a survey of US gastroenterologists. *Endosc Int Open* 2020;08:E1379–84.
13 **Schuitmaker L**, Drogt J, Benders M, *et al*. Physicians' required competencies in AI-assisted clinical settings: a systematic review. *Br Med Bull* 2025;153:ldae025.
14 **Obermeyer Z**, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
15 **Omiye JA**, Lester JC, Spichak S, *et al*. Large language models propagate race-based medicine. *NPJ Digit Med* 2023;6:195.
16 **Yang Y**, Liu X, Jin Q, *et al*. Unmasking and quantifying racial bias of large language models in medical report generation. *Commun Med* 2024;4:176.
17 **Hager P**, Jungmann F, Holland R, *et al*. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 2024;30:2613–22.
18 **Omar M**, Sorin V, Collins JD, *et al*. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Commun Med* 2025;5:330.