Current Research in Medical Sciences
p ISSN: 3116-0298

Student Research Committee
Babol University of Medical Sciences

Correspondence

From AI Literacy to Appropriate Reliance: An Unresolved Educational Question in Medical Education

Seyed Reza Hosseini[1]*, Alireza Abasikia[1]

[1] Student Research Committee, Babol University of Medical Sciences, Babol, Iran

Article Info.

Received: 15 May 2025
Accepted: 13 Dec 2025

*** Corresponding Author:**

Seyed Reza Hosseini
E-mail:
seyed265reza@gmail.com
ORCID: 0009-0001-6204-221X

Cite this article
Hosseini SR, Abasikia A. From AI Literacy to Appropriate Reliance: An Unresolved Educational Question in Medical Education. Curr Res Med Sci. Publication details pending.
DOI:10.22088/crms.9.2.81

Abstract

Generative artificial intelligence (AI) is increasingly used in medical education to explain concepts, discuss clinical cases, and provide feedback. Yet awareness of AI's limitations does not necessarily enable learners to evaluate individual outputs accurately. This correspondence distinguishes AI literacy and readiness from observed performance in judging AI-generated clinical information. Existing studies suggest that learners may struggle to identify inaccurate or misleading outputs, while findings on advice-taking and AI-supported teaching vary by task and outcome. General warnings and self-reported readiness scores alone cannot demonstrate reliable appraisal of a particular recommendation. Medical education therefore needs a clearer account of what constitutes context-sensitive evaluation of AI-generated information and how such judgement should be recognized.

Keywords: Medical education, generative artificial intelligence, AI literacy, appropriate reliance, clinical judgement, automation bias

Copyright © 2023 Babol University of Medical Sciences. Published by Babol University of Medical Sciences. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International license (https://creativecommons.org/licenses/by-nc/4.0/). Noncommercial uses of the work are permitted, provided the original work is properly cited.

To the Editor,

Generative artificial intelligence (AI) is increasingly used in medical education to explain concepts, discuss cases, and provide feedback. These opportunities also ask learners to judge whether a particular output is trustworthy in its clinical context. AI-literacy frameworks appropriately emphasize critical appraisal and awareness of system limitations [1, 2]. However, knowing that AI can make mistakes is not the same as recognizing a specific error, and model-level performance does not establish a learner's ability to evaluate individual recommendations.

Studies of AI literacy and readiness illustrate this distinction. In a two-centre study of 377 German medical students, self-reported critical appraisal and practical use were rated more highly than technical understanding; the study assessed literacy and attitudes rather than students' observed performance when checking clinical advice [2]. A Persian version of the Medical Artificial Intelligence Readiness Scale for Medical Students has also undergone psychometric evaluation [3]. Such instruments can characterize readiness and support curriculum planning, but their scores should not be treated as direct evidence of competent evaluation in practice.

Model performance is a separate issue. In an evaluation using Iranian pre-internship examinations, GPT-4 outperformed the other tested language models, with performance varying by model and language [4]. These findings describe examination performance under specific test conditions; they do not show that every output is correct or that learners can reliably identify errors in otherwise plausible responses.

Direct studies of output evaluation raise concerns. Waldock et al. asked 148 final-year medical students to assess GPT-3.5 responses to ten clinical questions, five of which were answered incorrectly. The median proportion of outputs correctly evaluated was 61%; when vague or inaccurate justifications were excluded, it was 56% [5]. The study highlights a practical difficulty, but it did not establish the effectiveness of a particular teaching intervention. In a different population, a study of general-practice trainees reported 55% mean accuracy in identifying hallucinated ChatGPT-4o responses [6]. As these participants were trainees rather than undergraduate medical students, the result is relevant context, not a direct estimate for medical students.

Evidence also cautions against describing learners as uniformly over-reliant on AI. In a study involving 40 medical students, responses shifted toward ChatGPT's answers whether those answers were correct or incorrect, and group discussion did not reliably correct AI-associated misconceptions [7]. By contrast, a randomized trial involving 186 fourth-year medical students found that a prominent warning that ChatGPT can make mistakes did not change diagnostic advice-taking (15.3% without the warning versus 15.9% with it). The investigators reported low weighting of AI advice in the task [8]. These results suggest that reliance depends on the task and context; neither habitual acceptance nor blanket rejection should be assumed.

Findings from AI-supported teaching are similarly difficult to interpret as evidence about calibrated reliance. In a randomized trial of 129 first-year medical students, ChatGPT-generated feedback and expert-written feedback produced no significant overall difference in clinical-reasoning scores immediately after teaching or at ten-day follow-up, although expert feedback performed better for more complex urinary-tract-infection cases [9]. A 2025 meta-analysis of 11 randomized trials involving 786 medical students likewise found no statistically significant pooled difference in knowledge acquisition between generative-AI-based and traditional teaching (standardized mean difference 0.27, 95% CI −0.31 to 0.85); the authors noted heterogeneity and methodological limitations [10]. These findings concern learning outcomes broadly and do not, by themselves, resolve how well learners judge the reliability of individual AI outputs.

The defensible gap is therefore not an absence of research on AI literacy, readiness, output evaluation, or AI-supported teaching. Rather, it remains unclear how reliably educational preparation translates into sound judgement about AI-generated information in context. This distinction matters because appropriate reliance is neither maximum trust nor maximum skepticism: it involves considering an AI response in light of the case, available evidence, and uncertainty. Earlier research on automation bias also shows that decision-support systems can contribute to error when verification is difficult, although much of that evidence predates generative AI [11].

Medical education should consequently distinguish self-reported familiarity from demonstrated judgement and avoid equating confidence, satisfaction, or general learning gains with reliable appraisal of AI-generated information. The aim should be informed, context-sensitive use not unquestioning acceptance, and not reflexive rejection. Whether current educational approaches consistently achieve that aim remains an important unanswered question.

Acknowledgment

Not applicable

Conflict of Interests

The authors declare no conflict of interest related to this study.

Author's contribution

All authors contributed to the conception and design of the study, data collection, data analysis, interpretation of the results, and preparation and revision of the manuscript. All authors reviewed and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Funding

None

Ethics statement

None

References

1. Gordon M, Daniel M, Ajiboye A, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024;46(4):446-470.
2. Laupichler MC, Aster A, Meyerheim M, Raupach T, Mergen M. Medical students' AI literacy and attitudes towards AI: a cross-sectional two-center study using pre-validated assessment instruments. BMC Med Educ. 2024;24:401.
3. Moodi Ghalibaf AA, Moghadasin M, Emadzadeh A, Mastour H. Psychometric properties of the Persian version of the Medical Artificial Intelligence Readiness Scale for Medical Students (MAIRS-MS). BMC Med Educ. 2023;23:577.
4. Zare S, Vafaeian S, Amini M, Farhadi K, Vali M, Golestani A, et al. Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams. Sci Rep. 2024;14:28456..
5. Waldock WJ, Lam G, Baptista A, Walls R, Sam AH. Which curriculum components do medical students find most helpful for evaluating AI outputs? BMC Med Educ. 2025;25:195.
6. Zhou J, Zhang J, Wan R, Cui X, Liu Q, Guo H, et al. Integrating AI into clinical education: evaluating general practice trainees' proficiency in distinguishing AI-generated hallucinations and impacting factors. BMC Med Educ. 2025;25:406.
7. Kirpalani A, Grimmer J, Wang PZT. Exploring medical student trust in generative artificial intelligence (ChatGPT) versus peers in team-based learning. Med Sci Educ. 2025;35(2):671-674.
8. Kıyak YS, Coşkun Ö, Budakoğlu İİ. 'ChatGPT can make mistakes' warnings fail: a randomized controlled trial. Med Educ. 2026;60(2):138-142
9. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. Postgrad Med J. 2025;101(1195):458-463.
10. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: a meta-analysis of randomized controlled trials. BMC Med Educ. 2025;25:1175.
11. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc.2017;24(2):423-431.