SHORT COMMUNICATION

Volume 18 No. 3 2026

DOI: 10.21315/eimj2026.18.3.18

ARTICLE INFO

Submitted: 23-11-2025
Accepted: 11-02-2026
Online: 30-09-2026

# Assessing the Authenticity of Artificial Intelligence Appraisals of Clinical Psychiatric Scenarios: A Preliminary Investigation

Marcus A. Henning[1], Zhaochu Geng[2], Christian U. Krägeloh[3], Lillian Ng[4], Yan Chen[1], Angel Li[2], Craig S. Webster[1]

[1]*Centre for Medical and Health Sciences Education, University of Auckland, NEW ZEALAND*

[2]*Faculty of Medical and Health Sciences, University of Auckland, NEW ZEALAND*

[3]*Faculty of Health and Environmental Sciences, Auckland University of Technology, NEW ZEALAND*

[4]*Department of Psychological Medicine, University of Auckland, NEW ZEALAND*

**To cite this article:** Henning MA, Geng Z, Krägeloh CU, Ng L, Chen Y, Li A, Webster CS. Assessing the authenticity of artificial intelligence appraisals of clinical psychiatric scenarios: a preliminary investigation. Education in Medicine Journal. 2026;18(3):265–274. https://doi.org/10.21315/eimj2026.18.3.18

**To link to this article:** https://doi.org/10.21315/eimj2026.18.3.18

## ABSTRACT

This short communication explores the utility of generative artificial intelligence (GenAI) systems in helping students prepare for clinical examinations. Four clinical scenarios from an objective structured clinical examination (OSCE) in psychiatry were presented to GPT-4o mini. This GenAI system was asked to respond to four questions related to the clinical psychiatric scenarios. Two questions were designed to be straightforward and contextually clear, whereas two incorporated irrelevant information. The GenAI responses demonstrated noticeable difficulty in accurately and authentically interpreting the scenarios, particularly when extraneous content was introduced. While the answers to the first two questions were reasonably accurate, the responses to the remaining two were unacceptable, highlighting the system's inability to decipher ambiguous or unclear input. Currently, GenAI systems appear to perform well at solving clinical cases when the data are straightforward and rule-based. However, GPT-4o mini demonstrated difficulty in interpreting scenarios complicated by emotional nuances or distracting elements.

**Keywords:** *OSCE, Psychiatry, Student learning, Generative artificial intelligence, GenAI*

## CORRESPONDING AUTHOR

Marcus A. Henning, Centre for Medical and Health Sciences Education, Faculty of Medical and Health Sciences, University of Auckland, Auckland 1023, New Zealand

Email: m.henning@auckland.ac.nz

## INTRODUCTION

The importance of clinically assessing the proficiencies of students in the health professions cannot be overstated. In any health professional education programme, assessing students' clinical capabilities ultimately aims to ensure that students have the skills, attitudes and knowledge to implement effective diagnostic processes and treatment interventions. In medical education, clinical assessment ensures that medical students responsibly adopt

© Malaysian Association of Education in Medicine and Health Sciences and Penerbit Universiti Sains Malaysia. 2026
This work is licensed under the terms of the Creative Commons Attribution (CC BY)
(http://creativecommons.org/licenses/by/4.0/).

professional attitudes as they progress through their medical training (1). Moreover, assessing clinical competence ensures that students have acquired adequate knowledge to practice effectively, efficiently, safely and responsibly. Assessing clinical competence also ensures that students cultivate a sense of social responsibility and cultural safety commensurate with their stage of training (1,2).

In medical education, the evaluation of clinical competency takes several forms. Common structured methods of assessing clinical competence include objective structured clinical examinations (OSCEs), single case-based examinations (SCBEs), direct observation of procedural skills (DOPS) and the mini-clinical evaluation exercise (Mini-CEX) (3,4). Some assessments are standardised, whereas others are context-specific; for example, the OSCE aims to present a standardised structure (4) for objectively assessing medical students' clinical competencies (3). The OSCE can be conducted in all medical disciplines and has been used in other disciplines, such as nursing and psychology (5–7). This paper focuses on the use of the OSCE in medical education, specifically in the evaluation of psychiatric competencies.

The advent of advanced technologies such as artificial intelligence (AI) has brought new challenges and opportunities in preparing students for clinical examinations and assessments of clinical competency. Vitands (8), a corporate communications specialist, has described how AI can be applied in clinical settings to support the assessment of clinical competency. The author explains that before the COVID-19 pandemic in the United States (9), the National Board of Medical Examiners was the key organisation overseeing licensing exams. Assessment in this context required that students take a computer-administered test on clinical reasoning as well as an in-person clinical skills assessment. The pandemic forced organisations to change their strategy, bringing more technologically advanced methods into play, creating the opportunity to use AI to support the assessment of clinical skills. AI software can harness natural language processing to aid in scoring students' proficiencies in diverse clinical domains, such as diagnosis, in real time. It can also assist in providing immediate, comprehensive feedback on performance.

The connection between AI and clinical assessment has emerged as a critical discourse in academia (10–12), with specific interest in generative AI (GenAI) systems (13,14). In direct relation to the present study, the application of GenAI to improve OSCEs has been explored. Mishra et al. (10) state in reference to the OSCE that GenAI can help reduce examiner bias, provide immediate and detailed feedback and simulate realistic clinical scenarios. Nonetheless, they note that GenAI may lack the emotional intelligence skills often required to detect subtle nuances in human communication. Misra and Surash (11) assert that the OSCE is time- and resource-intensive for medical organisations, making its implementation costly. They suggest that GenAI systems could practically assist educators in areas of preparation such as case development, standardised patient training, assessment methods, and grading rubrics (11). Still, they suggest that careful auditing is required when using such technologies, implying that they are augmentative rather than replacement systems.

Research is also burgeoning on the use of GenAI technologies by clinical students (15–17). In one study, Sami et al. (17) found that medical students had positive attitudes when considering the use of GenAI systems, with many seeing the technology as effective and credible, optimising their study time and providing current medical information. This discourse indicates a growing reliance on the use of GenAI systems to complete education-related tasks.

Clinical competence remains a critical skill requiring assessment. GenAI systems, such as Generative Pre-trained Transformer 4 omni mini (GPT-4o mini), which is a free version of GPT-4 developed by OpenAI (see, https://www.ibm.com/think/topics/gpt-4o), are increasingly heralded as a way to complete tasks more efficiently and effectively, resulting in cost and time savings. However, the subtle nature of human communication must be considered when using GenAI technologies, and greater caution is demanded when using GenAI in assessment areas with a strong element of emotional intelligence and characterised by known sex and racial biases (18). Therefore, this project asked the specific research question, "How does GPT-4o mini perform in answering questions related to clinical scenarios from a psychiatry-based OSCE?"

## METHODS

### Participants

Within the research team, one author was a vocationally trained psychiatrist. This researcher, clinician, and academic appraised a set of GPT-4o mini responses to four clinical scenarios. The psychiatrist on the team had worked as a consultant in public mental health services and as an OSCE tutor and examiner in medical education for 10 years.

### Procedure

A medical student research assistant searched the internet to find details on four clinical scenarios commonly used in psychiatry-based OSCEs. GPT-4o mini was asked to respond to four types of questions per scenario. We used the free GPT-4o mini version of ChatGPT – a large language model (LLM) developed by OpenAI (19) – which likely reflects the experience of a broad user base due to its greater accessibility. In addition to being free, GPT-4o mini enables users to access advanced reasoning and analysis capabilities when tackling difficult tasks, such as interpreting OSCE scenarios. Each question was developed to assess four conditions, namely (a) an original OSCE scenario with no modifications; (b) the same scenario with added context, for example, "You are a fourth-year medical student at the University of Auckland, sitting the objective structured clinical examination (OSCE)"; (c) the same scenario with irrelevant information added to the setting only; and (d) the same scenario with irrelevant information added to both the setting and the task. Therefore, conditions (a) and (b) provided focused prompts, whereas conditions (c) and (d) provided distracting prompts, thus stimulating potential prompts that may occur accidentally during the clinical exam. The subsequent appraisal of GPT-4o mini's competency was conducted three times; after each time, the previous response was deleted from the server. The four OSCE scenarios were presented to GPT-4o mini for a clinical appraisal of the questions being asked. Each question was labelled as a condition. Each OSCE scenario in all four conditions was appraised three times.

### Instrument

The OSCE has been used extensively to assess medical students' clinical skills (20). Typically, the OSCE involves students rotating around a series of clinical stations. At each station, a student responds to a series of questions about a particular clinical scenario. A clinical assessor at each station assesses the student's skill and records their performance score on

a grading sheet. In this study, GPT-4o mini served as the simulated student, and one of the authors served as the appraiser. The appraiser was aware they were assessing GPT-4o mini responses to questions about the OSCE scenarios.

Hodges et al. (20) conducted an early validation study of the OSCE in the psychiatric teaching context. They found that the OSCE had adequate construct and concurrent validity and that the stations were authentic representations of actual clinical encounters. The present study used four psychiatry-based clinical OSCE scenarios, sourced from oscesense.com, an educational resource website produced by UK doctors specifically for undergraduate medical students preparing for OSCEs. The site is widely used in medical education and offers peer-reviewed, structured scenarios designed to simulate real-life clinical situations. This study's four scenarios were written by Dr. Rishil Patel and Dr. Sara Pender and peer-reviewed by Dr. Claudius Ho and Dr. Benjamin Armstrong. All the contributors were qualified medical practitioners with first-hand experience of OSCEs and active roles in supporting medical education. The four scenarios were presented to the GPT-4o mini platform along with typical questions associated with the conditions 1 to 4. The sources of the scenarios with scoring criteria are available via the links as follows:

a. Scenario 1: https://www.oscesense.com/osce-practice-overdose-history
b. Scenario 2: https://www.oscesense.com/osce-practice-psych-bipolar
c. Scenario 3: https://www.oscesense.com/osce-practice-psych-anorexia-nervosa
d. Scenario 4: https://www.oscesense.com/osce-practice-psych-ptsd

### Data Analysis

Descriptive statistics were generated for the dataset, given that only one assessor evaluated the GenAI system's authenticity and accuracy when responding to four questions about four clinical scenarios. First, measures of central tendency and variability were computed across the four scenarios and four conditions (Table 1). Friedman's related-samples non-parametric test was used to determine whether all responses across conditions were the same. Second, the reliability of responses across the three response encounters was estimated using the intraclass correlation coefficient (ICC).

## RESULTS

### Measures of Central Tendency and Variability

Table 1 displays the mean value of each condition for each scenario; this report focuses on the condition variations, which are clearly indicated. The marks tend to decrease as the GPT-4o mini encounter moves from conditions 1 to 4. Figure 1 corroborates the visual inspection of measures across the four scenarios and shows the measures are lower in magnitude as the evaluation measures were recorded across the four scenarios. Because this was a test case using only one assessor, Friedman's related-samples two-way analysis of variance by ranks was employed, with the results across the four conditions indicating that the response distributions were not the same: $\chi^2(3) = 21.43$, $p < 0.001$.

**Table 1:** Mean and (standard deviation) of the three responses across four scenarios and four conditions

| Condition | Mean (Standard deviation) | | | |
|---|---|---|---|---|
| | **Scenario 1** | **Scenario 2** | **Scenario 3** | **Scenario 4** |
| 1 | 60.00 (15.00) | 28.33 (2.89) | 23.33 (5.774 | 20.00 (0.00) |
| 2 | 61.67 (15.28) | 40.00 (0.00) | 31.67 (12.58) | 21.67 (2.89) |
| 3 | 43.33 (17.56) | 35.00 (0.00) | 23.33 (2.89) | 13.33 (2.89) |
| 4 | 33.33 (5.77) | 28.33 (5.77) | 15.00 (0.00) | 13.33 (2.89) |

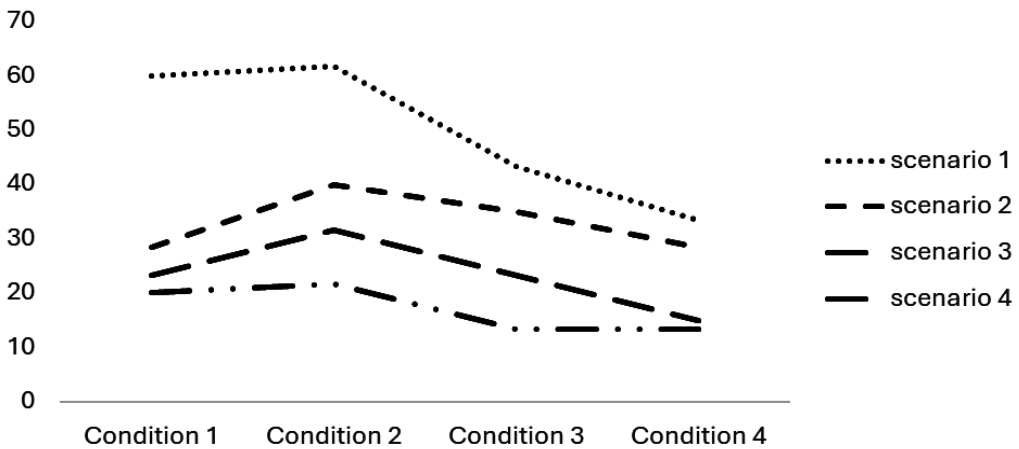



**Figure 1:** Mean OSCE value attributes to GPT-4o mini responses across the four conditions and according to each scenario.

### Reliability Across Response Encounters

Given the GPT-4o mini's competency was conducted three times, the reliability of the three response encounters was computed across all conditions and scenarios. Response 1 represents the first GPT-4o mini encounter, followed by responses 2 and 3. The ICC average statistic (0.89) indicates that how GPT-4o mini responded to each scenario across each condition demonstrated excellent reliability across the three response opportunities (21).

## DISCUSSION

This brief communication examined the authenticity and accuracy of GPT-4o mini's responses to questions about clinical scenarios from a psychiatry-based OSCE. The motivations for this study were manifold. First, given rising student interest in using GPT-4o mini to aid their clinical studies (17), we wished to assess GPT-4o mini's authenticity and accuracy if students use the technology to enhance their ability to answer clinical questions within a structured exam format (i.e., the OSCE). Second, we wanted to determine GPT-4o mini's credibility when given distracting prompts, which may inadvertently occur during clinical exams. Third, we were interested to see whether any trends emerged that could inform the future use of current GenAI technology. These trends were also interpreted to inform future advice to students who are motivated to use GPT-4o mini technology.

## Authenticity and Accuracy of AI During the OSCE

In psychiatry, an understanding and integration of biological, psychological and social perspectives is required to optimally respond to questions in standardised patient scenarios (22). Students also need a high degree of emotional intelligence to detect non-verbal cues and interpret nuances of human behaviour and interaction (23). The results of this study indicate that when the original OSCE scenario with no modifications was presented to GPT-4o mini, responses to scenarios 1 and 2 received higher scores than those to scenarios 3 and 4. Therefore, when provided with scenarios on the topics of overdose history and bipolar conditions, GPT-4o mini was able to generate acceptable responses. However, for scenarios on the topics of anorexia nervosa and post-traumatic stress disorder (PTSD), the responses were deemed less authentic and less accurate, reducing the GenAI's performance. Thus, the algorithms governing GPT-4o mini may discern certain conditions better than others, supporting the belief that the subtle nature of human communication may pose a challenge for GPT-4o mini, which likely requires further refinement (10). This suggests that certain standardised conditions are easier to resolve than ones with more linguistic subtlety or contradiction. Therefore, while GenAI systems are good at finding patterns and rules in data, they have difficulty when those patterns and rules are ambiguous. Unlike in real OSCE exams, however, we did not allow the AI to ask follow-up questions. GenAI systems can probably improve in several ways when interpreting complex cases, such as by exposing LLMs to examples that capture multifaceted nuances. Achieving this will likely require detailed, sophisticated feedback mechanisms and the use of wide-ranging datasets. Nonetheless, the findings of this preliminary study indicate that, at the current stage of AI development, the interface between clinicians and AI is in a state of transition, manifesting as an evolving process that ultimately aims to ensure optimal patient care. Ongoing collaboration among clinicians, medical educators and AI system developers is essential to reduce medical error and prevent subsequent performance issues.

## Adaptability of AI

This preliminary work examined a GenAI system's adaptability when faced with the same scenario with subtle changes: adding context, providing irrelevant information about the setting, and presenting irrelevant information about the setting and the task. To assess these changes, we added three further conditions beyond condition 1, the original OSCE scenario with no modifications. The results show that response accuracy decreased in conditions 3 and 4 but marginally improved in condition 2. Therefore, we surmised that when presented with confusing information, GPT-4o mini's responses became more blurred and inaccurate, indicating that it could not identify distracting information and focused only on data relevant to solving the task at hand (24–26).

Consistent with the findings of this study, Mahbub et al. (26) report identifiable limitations of LLMs, specifically GPT-4, which had problems in cross-referencing within a written document and demonstrated a limited understanding of specialised contexts. Mahbub et al. indicate that GPT-4 was accomplished in identifying incomplete requirements but exhibited problems when detecting inconsistencies and ambiguities. Given that clinical reasoning requires sophisticated management of conflicting solutions when solving complex cases, it is likely that more sophisticated software development is required to optimally develop solutions to such cases. This argument also applies to the notion of ambiguity, which likely occurs when different stakeholders perceive the solution differently.

## Future Use of Current GenAI Technology

GPT-4o mini is already becoming outmoded, with new versions and various other types of GenAI systems now available (27). He et al. (27) have identified several challenges in the use of GenAI systems. First, they exhibit poor transparency due to GenAI's inherent black box nature. Second, there is a trade-off between generalisation and performance, especially when a given model may perform better on one sample than another. This may partly explain the problems with maintaining high performance across all case scenarios in our study. Nonetheless, He et al. suggest that GenAI has shown great potential in clinical support, such as early warning systems or care assistance tools. Future GenAI tools that tackle these challenges will become paramount in health care.

Prather et al. (28) systematically reviewed the literature on current trends in GenAI, focusing on the computer education context; many of the points raised can also be applied to the health professional education context (13,14). In their review, Prather et al. (28) report that 36% of surveyed faculty and educators adopted GenAI systems, suggesting that future academics will increasingly adopt the technology rather than ban it. Several issues need to be resolved to make its use more meaningful and authentic. Importantly, educators need to be upskilled in the use of GenAI, and this upskilling needs to occur on a regular basis as the technology develops. This would entail regular communication between developers and educators so that the gap between education and development does not widen. This would ensure that, with increased knowledge and skills, trust can be guaranteed.

The specific findings of this study suggest cautionary considerations as well as merits when using GenAI systems (13). Many of the identified risks are consistent with misalignment across development, understanding, and usage. As shown in our study, inaccuracies likely occur related to the clinical assessment of student proficiency, and these must be acknowledged as we increasingly rely on GenAI in both health professional training and practice. Second, authenticity and integrity must be aligned so that the duty and responsibility for patient care are assured and ideally optimised. Third, bias in the ensuing scripts and information obtained must be audited and checked to ensure best practice is maintained (18). Lastly, as we become more reliant on GenAI, information storage becomes an issue, potentially affecting privacy and security (29). GenAI's potential to save time and resources in helping students prepare for a clinical assessment must be matched by caution in its use for real-world practice. There is no substitute for the ultimate test of assessing skills in a clinical setting, where students are face-to-face with a real patient.

## Strengths and Limitations

A strength of this study is its examination of the veracity and utility of GenAI – in this case, GPT-4o mini – in the clinical educational context. We focused on a clinical assessment used widely around the globe, the OSCE, a widely used clinical assessment, and considered how it could hinder or assist students in understanding clinical cases in the psychiatric context. We further contribute to the literature by considering the inter-rater reliability of the GPT-4o mini output and by assessing the accuracy and authenticity of responses through an experienced psychiatry OSCE assessor (one of the co-authors). A limitation is that this investigation was a preliminary study requiring further data to validate its findings. However, the procedure may inform a larger study that requires the engagement of more experienced OSCE assessors and the review of more case scenarios. This study's findings may also be applicable to studies examining how well diverse groups use GenAI-generated scenarios.

## CONCLUSION

All students in healthcare-related disciplines (e.g., medicine, nursing and psychology) need to be judiciously assessed to ensure optimal patient safety when they become proficient, registered clinicians. The OSCE is a commonly used clinical assessment that has demonstrated psychometric rigour. In recent years, GenAI systems, GPT-4o mini in particular, have been marketed as tools that can assist students in their learning and exam preparation. To make our investigation more focused and targeted, we considered how to appraise the accuracy and authenticity of GPT-4o mini by concentrating on psychiatric OSCEs in the medical education context. The findings of this small study suggest that GPT-4o mini can produce reasonably accurate and authentic responses when appraising unambiguous clinical cases, but its responses are not acceptable when irrelevant information is added to the questions. Therefore, we surmise that accuracy can be enhanced when contextual cues are provided, whereas inaccuracies become more prominent when irrelevant cues are provided. Therefore, the human skills often employed in psychiatry, such as emotional intelligence, are not easily understood by current GenAI systems.

## ACKNOWLEDGEMENTS

This project was partially supported by the University of Auckland Summer Research Scholarship (Ref no.: SUMRESCH 301).

## ETHICAL APPROVAL

This research did not involve human participants. Therefore, formal ethical approval was not required. All data were generated using GPT-4o mini. The consultant responsible for verifying the veracity of the GenAI responses is one of the co-authors. The authors declare no conflicts of interest.

## REFERENCES

1. Lindgren S, Argullos JLP, Millan JR. Assessment of clinical competence of medical students: future perspectives for Spanish faculties. Med Clin Pract. 2024;7(2):100424. https://doi.org/10.1016/j.mcpsp.2023.100424

2. Tervalon M. Components of culture in health for medical students' education. Acad Med. 2003;78(6):570–6. https://doi.org/10.1097/00001888-200306000-00005

3. Badyal DK, Sharma M. Internal assessment in new MBBS curriculum: methods and logistics. Int J Appl Basic Med Res. 2020;10(2):68–75. https://doi.org/10.4103/ijabmr.IJABMR_18_20

4. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357(9260):945–9. https://doi.org/10.1016/S0140-6736(00)04221-5

5. Mitchell ML, Henderson A, Groves M, Dalton M, Nulty D. The objective structured clinical examination (OSCE): optimising its value in the undergraduate nursing curriculum. Nurse Educ Today. 2009;29(4):398–404. https://doi.org/10.1016/j.nedt.2008.10.007

6. Olivares XG, de León DA. Objective structured clinical examination (OSCE) in psychology. In: Hosseini S, Beitzel B, Köhler M, Sanjakdar F, editors. Technological advancements and innovations in future of education. Singapore: Springer; 2024. p. 53.

7. Sheen J, McGillivray J, Gurtman C, Boyd L. Assessing the clinical competence of psychology students through objective structured clinical examinations (OSCEs): student and staff views. Aust Psychol. 2015;50(1):51–9. https://doi.org/10.1111/ap.12095

8. Vitands B. Artificial intelligence behind improved clinical skills assessment and teaching. Medical School News [cited 2025 Jun 24]. Available from: https://med.umn.edu/news/artificial-intelligence-behind-improved-clinical-skills-assessment-and-teaching

9. Ciotti M, Ciccozzi M, Terrinoni A, Jiang WC, Wang CB, Bernardini S. The COVID-19 pandemic. Crit Rev Clin Lab Sci. 2020;57(6):365–88. https://doi.org/10.1080/10408363.2020.1783198

10. Mishra GV, Luharia AA, Naqvi W, Sood A. Artificial intelligence in OSCE: innovations and implications for medical education assessment–a systematic review. In: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI). IEEE; 2024. p. 1–5. https://doi.org/10.1109/IDICAIEI61867.2024.10842789

11. Misra SM, Suresh S. Artificial intelligence and objective structured clinical examinations: using ChatGPT to revolutionise clinical skills assessment in medical education. J Med Educ Curric Dev. 2024;11:23821205241263475. https://doi.org/10.1177/23821205241263475

12. Soong TK, Ho CM. Artificial intelligence in medical OSCEs: reflections and future developments. Adv Med Educ Pract. 2021;12:167–73. https://doi.org/10.2147/AMEP.S287926

13. Janumpally R, Nanua S, Ngo A, Youens K. Generative artificial intelligence in graduate medical education. Front Med (Lausanne). 2025;11:1525604. https://doi.org/10.3389/fmed.2024.1525604

14. McCoy L, Ganesan N, Rajagopalan V, McKell D, Niño DF, Swaim MC. 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. https://doi.org/10.1177/23821205251339226

15. Mousavi Baigi SF, Sarbaz M, Ghaddaripouri K, Ghaddaripouri M, Mousavi AS, Kimiafar K. Attitudes, knowledge, and skills towards artificial intelligence among healthcare students: a systematic review. Health Sci Rep. 2023;6(3):e1138. https://doi.org/10.1002/hsr2.1138

16. Tangadulrat P, Sono S, Tangtrakulwanich B. Using ChatGPT for clinical practice and medical education: cross-sectional survey of medical students' and physicians' perceptions. JMIR Med Educ. 2023;9:e50658. https://doi.org/10.2196/50658

17. Sami A, Tanveer F, Sajwani K, Kiran N, Javed MA, Ozsahin DU, et al. Medical students' attitudes toward AI in education: perception, effectiveness, and its credibility. BMC Med Educ. 2025;25(1):82. https://doi.org/10.1186/s12909-025-06574-x

18. Webster CS, Jowsey T. Artificial intelligence must operate ethically in health care and not be prone to racist or sexist biases. Anesth Analg. 2025;140(5):1099–104. https://doi.org/10.1213/ANE.0000000000007131

19. OpenAI. ChatGPT: get answers, find inspiration, be more productive. 2025. Available from: https://openai.com/chatgpt/overview/

20. Hodges B, Regehr G, Hanson M, McNaughton N. Validation of an objective structured clinical examination in psychiatry. Acad Med. 1998;73(8):910–2. https://doi.org/10.1097/00001888-199808000-00021

21. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012

22. Hodges BD, Hollenberg E, McNaughton N, Hanson MD, Regehr G. The psychiatry OSCE: a 20-year retrospective. Acad Psychiatry. 2014;38(1):26–34. https://doi.org/10.1007/s40596-013-0008-8

23. Shafti SS. Association between emotional intelligence and objective structured examinations: a study on psychiatric residents. Educ Res Med Sci. 2017;6(1):23–8. https://doi.org/10.22110/erms.2017.51478

24. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4o mini on medical challenge problems. arXiv. 2023;arXiv:2303.13375. https://doi.org/10.48550/arXiv.2303.13375

25. Arkoudas K. GPT-4o mini can't reason. arXiv. 2023;arXiv:2308.03762. https://doi.org/10.48550/arXiv.2308.03762

26. Mahbub T, Dghaym D, Shankarnarayanan A, Syed T, Shapsough S, Zualkernan I. Can GPT-4 aid in detecting ambiguities, inconsistencies, and incompleteness in requirements analysis? A comprehensive case study. IEEE Access. 2024;12:12345–56. https://doi.org/10.1109/ACCESS.2024.3350000

27. He R, Sarwal V, Qiu X, Zhuang Y, Zhang L, Liu Y, et al. Generative AI models in time-varying biomedical data: scoping review. J Med Internet Res. 2025;27:e59792. https://doi.org/10.2196/59792

28. Prather J, Leinonen J, Kiesler N, Gorson Benario J, Lau S, MacNeil S, et al. Beyond the hype: a comprehensive review of current trends in generative AI research, teaching practices, and tools. In: 2024 Working Group Reports on Innovation and Technology in Computer Science Education. New York: Association for Computing Machinery; 2025. p. 300–38. https://doi.org/10.1145/3689187.3709614

29. Ng BY, Li J, Tong X, Ye K, Yenne G, Chandrasekaran V, et al. Analyzing security and privacy challenges in generative AI usage guidelines for higher education. arXiv. 2025; arXiv:2506.20463. https://doi.org/10.48550/arXiv.2506.20463