THE EUROPEAN SOCIETY OF MEDICINE
Medical Research Archives, Volume 14 Issue 3

RESEARCH ARTICLE

Generative Artificial Intelligence versus Faculty: A Comparative Study of Narrative Feedback on Medical Students' Written Mental Status Exams

Elle S. Cleaves[1], David C. Belmonte[2], Sorana Raiciulescu[1], Joshua R. Duncan[1]

[1]Uniformed Services University of the Health Sciences, Bethesda, Maryland, United States of America
[2]University of Michigan, Ann Arbor, Michigan, United States of America

OPEN ACCESS

PUBLISHED
31 March 2026

CITATION
Cleaves, E.S., et al., 2026. Generative Artificial Intelligence versus Faculty: A Comparative Study of Narrative Feedback on Medical Students' Written Mental Status Exams. Medical Research Archives, [online] 14(3).

ISSN
2375-1924

ABSTRACT

Objective: The authors evaluated narrative feedback generated by ChatGPT and faculty members for medical students' Mental Status Exam write-ups. The study compared feedback quality and usefulness and assessed whether students and academic psychiatrists could identify the feedback source.

Methods: Medical students (N=164) wrote Mental Status Exams and received blinded feedback from either a faculty member (for low-scoring write-ups, n=43) or ChatGPT (for high-scoring write-ups, n=121). Students rated the feedback's quality and usefulness and guessed its origin. Three academic psychiatrists also conducted a blinded evaluation, rating both feedback types for the low-scoring Mental Status Exams, choosing the superior version, and guessing the source.

Results: Students rated Artificial Intelligence-generated feedback quality significantly higher than faculty feedback (mean=4.22 vs. 3.5). Academic psychiatrists preferred the Artificial Intelligence-generated feedback in 93% of cases. Only 29% of students receiving Artificial Intelligence-generated feedback correctly identified its source. Psychiatrists correctly identified Artificial Intelligence-generated feedback only 23% of the time and misattributed faculty feedback as Artificial Intelligence-generated 71% of the time.

Conclusions: Artificial Intelligence-generated feedback was perceived as high-quality by students and preferred by expert raters. The difficulty in distinguishing Artificial Intelligence-generated from faculty feedback suggests generative Artificial Intelligence can produce feedback comparable or superior to human experts, offering a scalable tool to support medical education and reduce faculty workload.

Keywords: Artificial Intelligence, Medical Education, Feedback, Mental Status Examination, Psychiatry

Introduction

High-quality feedback is essential in medical education for the development of clinical skills, as it enables learners to identify strengths and areas for improvement, promotes self-reflection, and supports progression toward competency. In psychiatric training, the Mental Status Examination (MSE) is a fundamental competency,[1] requiring nuanced clinical reasoning and communication skills. Effective feedback on MSE performance is critical to ensure trainees achieve proficiency in this core skill.[1,2]

In competency-based medical education, timely, specific, and actionable feedback is a cornerstone for learner development.[1,2] However, faculty often struggle with providing consistent, individualized narrative feedback due to increasing learner-to-faculty ratios and clinical demands. There is also a need for standardization to ensure consistency and fairness across evaluators, yet faculty feedback can be highly variable. Providing personalized, actionable feedback at scale is difficult, often resulting in delayed or generic responses that may not optimally support learner development.[3]

Generative AI (AI) can improve personalized learning and enhance learning by providing feedback.[4,5] It can also potentially increase the amount, depth, and quality of feedback provided to learners.[6] As a potential solution to the challenges of faculty-led feedback, generative AI offers the ability to deliver immediate, scalable, and standardized feedback on clinical documentation and simulated patient encounters, including the MSE. Generative AI has demonstrated the ability to assess learner submissions with consistency comparable to faculty, while dramatically reducing faculty workload and enabling timely, individualized feedback. Recent systematic reviews highlight the potential of generative artificial intelligence to address these gaps by offering scalable, personalized feedback that can match or exceed human-generated content in certain contexts.[7,8] Studies have shown generative artificial intelligence feedback improves clinical reasoning skills comparably to expert feedback in randomized trials[9] and demonstrates scoring consistency with faculty while reducing workload.[3] However, concerns remain regarding the accuracy, contextual appropriateness, and potential biases of AI-generated feedback.[3,10–12]

A key gap in the current literature is the lack of direct comparative studies evaluating the quality, usefulness, and learner perception of AI-generated feedback versus experienced faculty feedback, particularly in psychiatric education and MSE training. Several articles have been written about AI in psychiatry education in general but none have been written about narrative feedback and MSE training.[13–15] While preliminary evidence supports the complementary role of AI, rigorous research is needed to determine its effectiveness and acceptability relative to traditional faculty feedback in this context.[10,16,17] Therefore, this study evaluated the ability of ChatGPT to provide narrative feedback on medical students' Mental Status Exam (MSE) write-ups, compared the quality and usefulness of ChatGPT's feedback to a faculty member's feedback, and determined whether psychiatrists and medical students could correctly identify the source of the feedback.

Methods

Medical students were tasked with writing a MSE after watching a video of a patient interview, having been previously taught a 10-point MSE framework. Narrative feedback ("Recommendations") was generated for 43 written MSEs that scored 7 or below out of 10 by a faculty member (a psychiatrist). For the AI-generated feedback, a customized ChatGPT interface using OpenAI's GPT-4o model, "MSE Grader",[18] was created to generate narrative feedback on deidentified written MSEs. Both the faculty member and ChatGPT were instructed to "provide 2-6 sentences with recommendations for how the student can improve documentation of their mental status exam".

Students were blinded to the source of their feedback, meaning they did not know if it came

from a faculty member or ChatGPT. Students with low scores (7 or below) received feedback from a faculty member while high-scoring students (8-10) received feedback from ChatGPT. Following receipt of feedback, students were surveyed to rate the quality of the feedback on a 5-point Likert scale (Very Poor, Poor, Acceptable, Good, or Very Good)[19] and its usefulness (low, moderate, or high). Usefulness ratings were defined as: Low (minimal specific information, often vague), Moderate (uses terms from rubric with minimal advice), or High (gives examples, helps understanding, reinforces strengths, gives constructive criticism).[20] Students were also asked to guess whether the feedback was from generative AI or a faculty member. Figure 1 depicts the flow chart of the process.

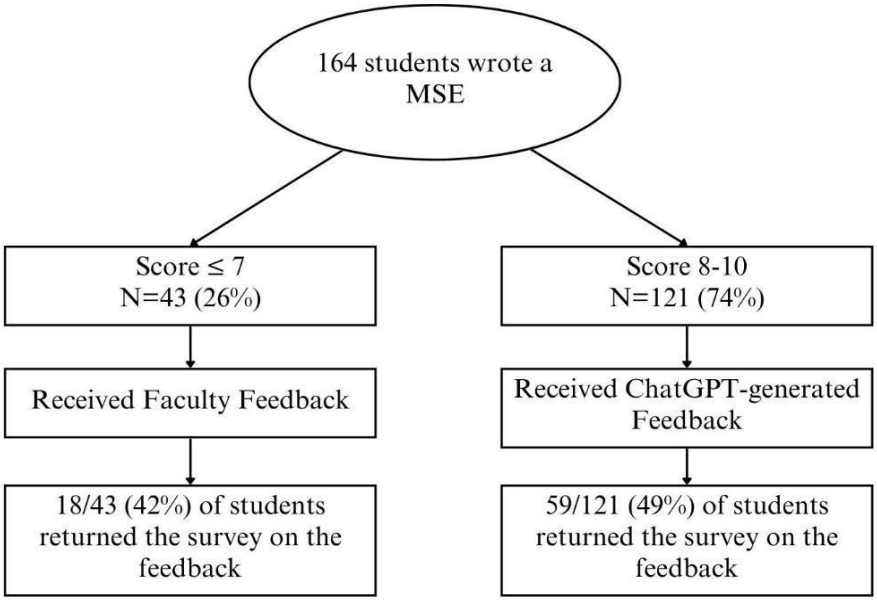

Figure 1: Medical students were given feedback on their mental status exam (MSE) write-ups and surveyed on the quality and usefulness of the feedback.

In a separate evaluation, three academic psychiatrists were shown the 43 low-scoring students' written MSEs each accompanied by two versions of narrative feedback: one written by ChatGPT and one by a faculty member. These raters were also blinded to the source of the feedback. The order of presentation for the AI and faculty feedback was randomized to prevent ordering effects. They rated the quality of each feedback version on the same 5-point Likert scale[19] and rated the usefulness as low, moderate, or high, using the specified criteria.[20] Figure 2 depicts the flow chart of this process. Additionally, the three academic psychiatrists were asked to choose which feedback was better and whether they thought it was generated by ChatGPT.

All analyses were performed using IBM SPSS Statistics for Windows, version 29 (IBM Corp., Armonk, NY, USA). Descriptive statistics summarized response rates, means (with standard deviations), and proportions. Student quality ratings (continuous 5-point Likert scores) between generative artificial intelligence and faculty feedback groups were compared using an independent-samples t-test with Welch's correction for unequal variances; effect size was calculated as Cohen's d. Usefulness ratings (categorized as moderate/high versus low) were summarized descriptively as percentages. Inter-rater reliability among the three psychiatrist raters was assessed using Fleiss' kappa (for overall multi-rater agreement) and Cohen's weighted kappa (for pairwise comparisons) on quality and usefulness ratings. Interpretation followed standard guidelines (e.g., <0.20 poor, 0.21-0.40 fair). Source-identification percentages were reported descriptively. Statistical significance was set at $p < 0.05$ (two-tailed). The Uniformed Services University Human Research Protections Program determined this study exempt (Protocol DBS.2024.768).

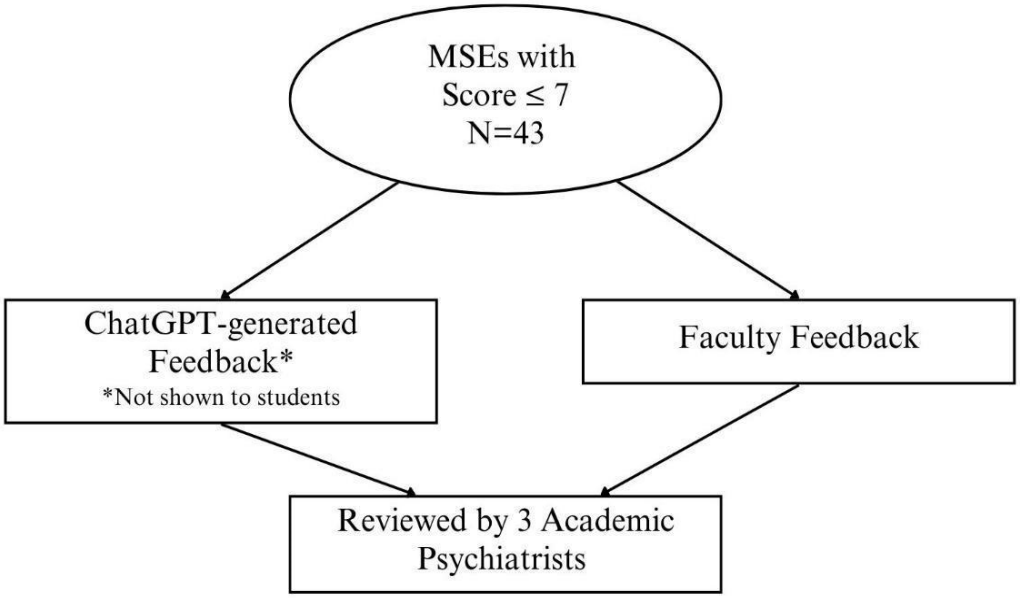

Figure 2: Three academic psychiatrists reviewed ChatGPT-generated feedback and faculty feedback on low-scoring written Mental Status Exams (MSEs).

Results

The study yielded distinct findings regarding perception and actual source of feedback among both students and academic psychiatrists. Of 164 eligible students, 43 scored ≤ 7 and received faculty feedback, while 121 students scored >7 and received AI feedback. The post-feedback survey was completed by 18 of the 43 students (42%) in the faculty feedback group and 59 of the 121 students (49%) in the AI feedback group.

Student ratings of feedback quality were significantly higher for the AI-generated group (mean=4.22, SD=0.70) compared to the faculty-generated group (mean=3.50, SD=0.86). This difference was statistically significant and represented a large effect size ($t(22.43)=2.89$, $p=0.008$, $d=0.78$). Similarly, students found the AI feedback more useful, with 95% (56 of 59) rating its usefulness as moderate or high, compared to 78% (14 of 18) for faculty feedback.

Among the three academic psychiatrists evaluating feedback for the 43 low-scoring MSEs, inter-rater agreement on the quality of narrative feedback was low with Cohen's Weighted Kappa ranging from 0.164 to 0.390. Agreement on the usefulness of feedback was also low (Kappa = 0.2). Despite the low agreement on specific ratings, the psychiatrists overwhelmingly preferred AI-generated feedback, selecting it as the better version in 93% of cases (see Figure 3).

Table 1 shows examples of faculty-generated feedback and ChatGPT-generated feedback for the same mental status exam write-ups. Both students and faculty raters had difficulty correctly identifying the feedback source. When asked to identify the source of their feedback, the majority of students (71%) believed their feedback originated from a faculty member. Specifically, only 29% of students who received AI-generated feedback correctly guessed that it was AI-generated (see Figure 4). Conversely, 72% of students who received faculty-generated feedback correctly identified it as such.

Furthermore, the three academic psychiatrists correctly identified AI-generated feedback only 23% of the time (see Figure 5). Remarkably, when the feedback was faculty-generated, the raters incorrectly thought it was AI-generated 71% of the time.

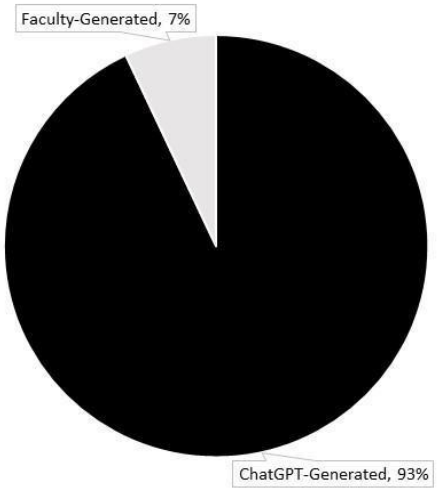

Figure 3: Faculty preferred ChatGPT-generated feedback 93% of the time.

Table 1: An example of faculty-generated feedback and ChatGPT-generated feedback for the same Mental Status Exam (MSE) write-up

Faculty-Generated Feedback	ChatGPT-Generated Feedback
Please keep things such as Behavior, which is observation non-interpretive. Please use the 5 aspects of affect (congruent, aspect, lability, range, and intensity).	Avoid describing cognition as impaired without specific tests or clear clinical evidence. State "not tested" or describe observable features suggesting cognition levels. Clarify the Thought Process by distinguishing the patient's ability to maintain a coherent dialogue versus the influence of hallucinations on logical thinking. Describe logical structures or breaks in logic more precisely. Continue providing detailed observations, ensuring comprehensive assessment in each section for accuracy and usefulness of the mental status exam.
Please read article provided for review of MSE. Multiple sections were not correctly addressed, including how mood, affect, and cognition are judged.	Insight is marked as absent and judgment as poor, which accurately reflects the patient's condition but could benefit from more detailed examples of how this impacts her daily decision-making and safety. Thought Process needs clearer documentation of the illogical elements and how they relate to her beliefs about military involvement in her life.
Please review the aspects of affect and safety and why they are important in a psychiatric patient. Please review why insight and judgement are important in diagnosis and also correct disposition of psychiatric patients.	The Thought Process is noted as coherent yet illogical and circumstantial. Improve the assessment by providing specific examples of circumstantial thought patterns and explaining how these affect the patient's ability to communicate effectively. This detail will aid in understanding the impact of mental health on her daily functioning.

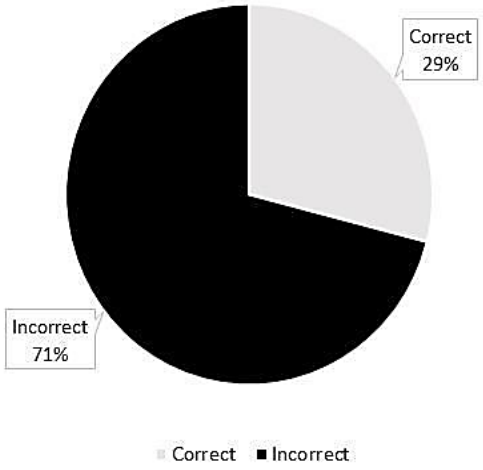

Figure 4: Only 29% of students guessed correctly that feedback was AI-generated

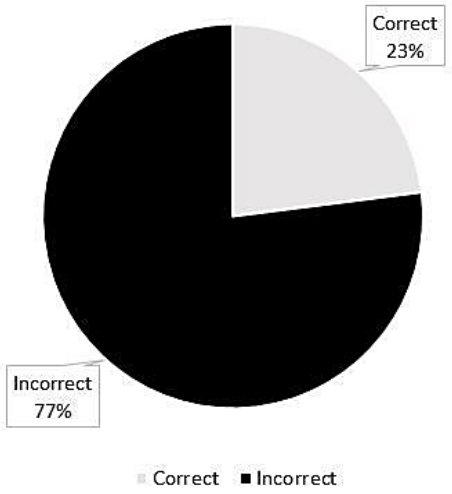

Figure 5: The faculty raters correctly identified AI-generated feedback 23% of the time

Discussion

This study addresses a critical gap by directly comparing generative artificial intelligence (ChatGPT-4o) narrative feedback to faculty feedback on medical students' written mental status exams. Students rated generative artificial intelligence feedback significantly higher in quality and usefulness, while blinded expert psychiatrists preferred it in 93% of cases despite low inter-rater agreement on specific scores. These findings align with and extend the growing body of evidence on generative artificial intelligence in health professions education. A recent systematic review found generative artificial intelligence frequently supports practice, inquiry, and production tasks (including feedback cycles), offering efficiency and personalization advantages, though human interaction remains valued for depth.[8] A 2025 meta-analysis of ChatGPT applications demonstrated large positive effects on learning performance (Hedges' g=0.867) and moderate gains in learning perception and higher-order thinking, particularly when used as an intelligent tutor providing personalized guidance.[21]

The overwhelming expert preference for generative artificial intelligence feedback (despite low kappa values indicating subjectivity in human rating) mirrors patterns in comparative studies. For instance, large language models showed scoring consistency comparable to faculty with dramatic workload reductions,[3] and randomized trials found ChatGPT

feedback on clinical reasoning as effective as expert input.[9] The examples in Table 1 illustrate how generative artificial intelligence often provides more structured, specific, and constructive recommendations—qualities that likely drove both student ratings and expert selection. Low inter-rater agreement (0.164-0.390 for quality) highlights a well-documented challenge in narrative evaluation; generative artificial intelligence offers standardization that may mitigate such variability.

These results contrast somewhat with one recent comparative study in which human tutor feedback was rated higher in certain dimensions.[16] Differences may stem from prompt engineering, model version (GPT-4o), or the specific task (Mental Status Examination documentation). In psychiatry education, where nuanced observation is key, generative artificial intelligence's consistency appears particularly advantageous. Students' difficulty distinguishing sources (only 29% correct for generative artificial intelligence) and psychiatrists' frequent misattribution further support that generative artificial intelligence can produce feedback indistinguishable from—or superior to—expert human output.

However, the study has limitations. The non-randomized design, which resulted in unequal group sizes based on initial student performance, may introduce selection bias. Additionally, the sample size was limited, which affects the generalizability of these findings to a broader population. Furthermore, one faculty member wrote all the narrative feedback, so the results may not be generalizable to all faculty members. The study also noted a low response rate from students on the survey, which could impact the representativeness of the student feedback data. Another limitation was the low inter-rater agreement among the academic psychiatrists, which suggests a high degree of subjectivity in evaluating feedback. However, this low statistical agreement stands in stark contrast to the near-unanimous preference (93%) the same raters showed for the AI-generated feedback. This suggests that while faculty may disagree on the specific numerical 'quality' of feedback, they can overwhelmingly recognize a superior product when they see it, which in this case was consistently generated by the AI. This discrepancy was particularly evident in the quality ratings, where the variability in how individual raters used the 5-point scale likely contributed to the low kappa scores. This finding highlights the inherent challenge of achieving standardized evaluation among human experts, a problem that AI-driven assessment may help to mitigate. Future analyses could consider collapsing these rating categories to a binary scale to potentially improve statistical agreement. A future study could examine the reasons why faculty members preferred one feedback instead of the other. A follow-up study could also ask students to compare faculty-generated feedback and AI-generated feedback and determine which feedback that students prefer and their reasons why.

Conclusion

Generative artificial intelligence, exemplified by ChatGPT, shows strong potential for delivering high-quality narrative feedback in medical education. In this study, generative artificial intelligence-generated feedback on Mental Status Examination write-ups was rated higher by students and overwhelmingly preferred by academic psychiatrists compared to faculty feedback. The difficulty in distinguishing generative artificial intelligence from human feedback underscores its readiness for integration into educational workflows. By augmenting traditional methods, generative artificial intelligence can help mitigate faculty workload pressures while maintaining or improving feedback quality. Further research is needed to evaluate long-term educational outcomes, optimize prompts, ensure equity, and explore applications across other clinical documentation and specialties.

Acknowledgments:

Derrick Hamaoka, Christina LaCroix, Mary Steinmann, Sean Wilkes

Data availability:

The data that support the findings of this study are not openly available due to reasons of privacy and are available from the corresponding author upon reasonable request. Data are located in controlled access data storage per institutional policy.

Declarations:

On behalf of all authors, the corresponding author states that there is no conflict of interest. The opinions and assertions expressed herein are those of the author(s) and do not reflect the official policy or position of the Uniformed Services University of the Health Sciences or the Department of Defense.

Ethical Approval:

The Uniformed Services University Human Research Protections Program determined this study exempt (Protocol DBS.2024.768).

Funding:

Not Applicable.

References:

1. Amonoo HL, Longley RM, Robinson DM. Giving Feedback. *Psychiatric Clinics of North America*. 2021; 44(2):237-247. doi:10.1016/j.psc.2020.12.006

2. Lee GB, Chiu AM. Assessment and feedback methods in competency-based medical education. *Annals of Allergy, Asthma and Immunology*. 2022; 128(3):256-262. doi:10.1016/j.anai.2021.12.010

3. Sreedhar R, Chang L, Gangopadhyaya A, et al. Comparing Scoring Consistency of Large Language Models with Faculty for Formative Assessments in Medical Education. *J Gen Intern Med*. 2025;40(1): 127-134. doi:10.1007/s11606-024-09050-9

4. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Switzerland)*. 2023;11(6). doi:10.3390/healthcare11060887

5. Lee H. The rise of ChatGPT: Exploring its potential in medical education. *Anat Sci Educ*. 2024;17(5): 926-931. doi:10.1002/ase.2270

6. Boscardin CK, Gin B, Golde PB, Hauer KE. ChatGPT and Generative Artificial Intelligence for Medical Education: Potential Impact and Opportunity. *Academic Medicine*. 2024;99(1):22-27. doi:10.1097/ACM.0000000000005439

7. Natesan S, Jordan J, Sheng A, et al. Feedback in Medical Education: An Evidence-based Guide to Best Practices from the Council of Residency Directors in Emergency Medicine. *Western Journal of Emergency Medicine*. 2023;24(3):479-494. doi:10.5811/westjem.56544

8. Pham TD, Karunaratne N, Exintaris B, et al. The impact of generative AI on health professional education: A systematic review in the context of student learning. *Med Educ*. 2025;59(12):1280-1289. doi:10.1111/medu.15746

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. doi:10.1093/postmj/qgae170

10. Lee QY, Chen M, Ong CW, Ho CSH. The role of generative artificial intelligence in psychiatric education– a scoping review. *BMC Med Educ*. 2025;25(1). doi:10.1186/s12909-025-07026-9

11. Verghese BG, Iyer C, Borse T, Cooper S, White J, Sheehy R. Modern artificial intelligence and large language models in graduate medical education: a scoping review of attitudes, applications & practice. *BMC Med Educ*. 2025;25(1). doi:10.1186/s12909-025-07321-5

12. Janumpally R, Nanua S, Ngo A, Youens K. Generative artificial intelligence in graduate medical education. *Front Med (Lausanne)*. 2024; 11. doi:10.3389/fmed.2024.1525604

13. Pang HYM, Meshkat S, Teferra BG, et al. Opportunities and Barriers of Generative Artificial Intelligence in the Training of Psychiatrists: A Competencies-Based Perspective. *Academic Psychiatry*. 2025;49(1):25-30. doi:10.1007/s40596-024-02087-2

14. Torous J, Greenberg W. Large Language Models and Artificial Intelligence in Psychiatry Medical Education: Augmenting But Not Replacing Best Practices. *Academic Psychiatry*. 2025;49(1):22-24. doi:10.1007/s40596-024-01996-6

15. King DR, Liu HY, Brenner AM. Academic Psychiatry in the Age of Artificial Intelligence. *Academic Psychiatry. Springer Science and Business Media Deutschland GmbH*. 2025;49(1):1-4. doi:10.1007/s40596-025-02112-y

16. Ali M, Harbieh I, Haider KH. Bytes versus brains: A comparative study of AI-generated feedback and human tutor feedback in medical education. *Med Teach*. 2026;48(1):131-141. doi:10.1080/0142159X.2025.2519639

17. 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. doi:10.1080/0142159X.2024.2314198

18. Duncan J. MSE Grader. GPT-4o. OpenAI; 2024. Accessed May 20, 2024. https://chat.openai.com/g/g-xk9bqKrvg-mse-grader.

19. Ayers JW, Poliak A, Dredze M, et al. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

20. Kelly MS, Mooney CJ, Rosati JF, Braun MK, Stone RT. Education Research: The Narrative Evaluation Quality Instrument: Development of a tool to assess the assessor. *Neurology*. 2020;94(2):91-95. doi:10.1212/WNL.0000000000008794

21. Wang J, Fan W. The effect of ChatGPT on students′ learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanit Soc Sci Commun*. 2025;12(1). doi:10.1057/s41599-025-04787-y