

An Analytical Framework for Evaluating Generative Artificial Intelligence and Faculty-Derived Narrative Feedback in Medical Students' Written Mental Status Examinations

Dr. Reza Mahmoudi

Department of Computer Engineering,
Sharif University of Technology, Tehran, Iran

Dr. Fatemeh Hosseini

Faculty of Information Technology,
Amirkabir University of Technology, Tehran, Iran

VOLUME 03 ISSUE 07 (2026)

Published Date: 01 July 2026 // Page no.: - 01-05

ABSTRACT

The rapid integration of generative artificial intelligence (AI) into medical education has introduced transformative opportunities in assessment and feedback systems. Narrative feedback, particularly in written Mental Status Examinations (MSEs), plays a crucial role in shaping clinical reasoning, reflective practice, and competency development among medical students. However, the comparative effectiveness of AI-generated feedback versus faculty-derived feedback remains insufficiently explored. This study proposes a structured analytical framework to evaluate the qualitative, pedagogical, and cognitive dimensions of feedback generated by generative AI systems and human faculty.

Adopting an experimental research design, the study integrates multi-dimensional evaluation metrics including linguistic quality, clinical accuracy, pedagogical relevance, and learner impact. Drawing upon existing literature in AI-assisted medical education and feedback theory, the framework operationalizes both quantitative scoring mechanisms and qualitative thematic analysis. The findings suggest that while generative AI demonstrates high consistency, scalability, and structural coherence, faculty feedback retains superiority in contextual depth, clinical nuance, and adaptive mentorship. The study highlights a hybrid augmentation model as a viable pathway for optimizing feedback systems.

The research contributes to the growing discourse on AI-human collaboration in medical education by providing a validated evaluative architecture. It also identifies critical limitations related to bias, interpretability, and pedagogical alignment. Future research directions include longitudinal impact assessment and integration of adaptive learning systems.

Keywords: Generative Artificial Intelligence, Medical Education, Narrative Feedback, Mental Status Examination, Clinical Assessment, AI-Human Comparison, Feedback Framework, Educational Analytics, Psychiatry Training.

1. INTRODUCTION

Extravasation The increasing complexity of medical education necessitates robust assessment mechanisms that not only evaluate knowledge but also enhance clinical reasoning and reflective competence. Narrative feedback, particularly in psychiatry-focused written Mental Status Examinations (MSEs), serves as a critical pedagogical tool for fostering diagnostic reasoning and communication skills. Traditional feedback systems rely heavily on faculty expertise, which, although rich in contextual insight, are often constrained by time, scalability, and inter-rater variability (Amonoo et al., 2021).

The emergence of generative artificial intelligence has introduced new paradigms in automated feedback generation. Large language models have demonstrated

the capacity to produce structured, coherent, and contextually relevant feedback at scale, thereby addressing some limitations of traditional systems (Boscardin et al., 2024; Lee H., 2024). However, the pedagogical effectiveness and clinical reliability of such AI-generated feedback remain contested, particularly in nuanced domains such as psychiatric evaluation.

Recent studies have highlighted the comparative potential of AI and human feedback systems. For instance, Ali et al. (2026) demonstrate that AI-generated feedback can match human tutors in structural clarity but may lack interpretive depth. Similarly, Sreedhar et al. (2025) emphasize the consistency of AI scoring mechanisms, while Çiçek et al. (2025) identify differences in learning outcomes when comparing AI and expert feedback.

Despite these advancements, there is a lack of a unified analytical framework that systematically evaluates both AI-generated and faculty-derived narrative feedback across multiple dimensions. This research addresses this gap by proposing a comprehensive evaluation model tailored to written MSE assessments.

The objectives of this study are threefold: (1) to develop a multi-dimensional analytical framework for feedback evaluation, (2) to empirically compare AI-generated and faculty feedback using this framework, and (3) to derive implications for integrating AI into medical education. The scope is limited to written MSEs in undergraduate medical training, with a focus on psychiatry education.

2. LITERATURE REVIEW

The integration of artificial intelligence into medical education has been extensively explored, particularly in the context of assessment and feedback. Gordon et al. (2024) provide a comprehensive scoping review highlighting the increasing adoption of AI tools for personalized learning and automated evaluation. Their findings emphasize the potential of AI to enhance efficiency while raising concerns regarding validity and ethical considerations.

Generative AI systems, such as large language models, have demonstrated significant capabilities in producing educational content and feedback. Boscardin et al. (2024) argue that AI can transform medical education by enabling scalable and standardized feedback systems. Similarly, Lee H. (2024) discusses the transformative role of ChatGPT in enhancing student engagement and learning outcomes.

However, the effectiveness of AI-generated feedback remains a subject of debate. Ali et al. (2026) provide empirical evidence comparing AI-generated and human tutor feedback, revealing that while AI excels in consistency and clarity, it often lacks the nuanced understanding required for complex clinical reasoning. This finding aligns with the observations of Torous and Greenberg (2025), who caution against over-reliance on AI in psychiatry education.

The role of feedback in medical education has been well-established in the literature. Amonoo et al. (2021) and Natesan et al. (2023) emphasize that effective feedback must be timely, specific, and actionable to facilitate learning. Lee and Chiu (2022) further highlight the importance of aligning feedback with competency-based education frameworks.

Recent studies have also explored the impact of AI on learning outcomes. Çiçek et al. (2025) conducted a randomized controlled trial demonstrating that expert feedback leads to better clinical reasoning compared to

AI-generated feedback. Conversely, Sreedhar et al. (2025) found that AI models exhibit high scoring consistency, suggesting their utility in formative assessments.

Pham et al. (2025) and Sallam (2023) provide systematic reviews indicating both opportunities and challenges associated with AI in healthcare education. While AI offers scalability and personalization, concerns regarding bias, interpretability, and ethical use persist.

Despite these contributions, there remains a lack of a structured framework for evaluating narrative feedback in medical education. Existing studies primarily focus on either quantitative metrics or qualitative insights, without integrating both dimensions into a unified model. This study addresses this gap by proposing a comprehensive analytical framework.

3. METHODOLOGY

3.1 Research Design

This study adopts an experimental research design to evaluate and compare generative AI-generated feedback with faculty-derived narrative feedback in medical students' written Mental Status Examinations. The design integrates both quantitative and qualitative analytical approaches to ensure a comprehensive evaluation of feedback quality and effectiveness.

The experimental framework is grounded in educational assessment theory and AI evaluation models, incorporating multi-dimensional metrics to capture the complexity of feedback processes. The approach is informed by prior comparative studies in medical education, particularly the work of Ali et al. (2026), which highlights the need for structured evaluation mechanisms in AI-human feedback comparisons.

3.2 Conceptual Framework

The proposed analytical framework consists of four primary evaluation dimensions:

1. Linguistic Quality – coherence, clarity, grammar, and readability
2. Clinical Accuracy – correctness of psychiatric interpretations and diagnostic reasoning
3. Pedagogical Relevance – alignment with learning objectives and competency frameworks
4. Cognitive Impact – ability to promote critical thinking and reflective learning

Each dimension is operationalized using specific indicators derived from existing literature on feedback effectiveness (Amonoo et al., 2021; Natesan et al., 2023).

3.3 Data Collection and Sampling

A representative sample of written MSE responses from medical students is selected. Each response is evaluated using two feedback sources:

- Generative AI Feedback: Generated using a large language model-based system (Duncan, 2024)
- Faculty Feedback: Provided by experienced psychiatry educators

The sampling strategy ensures diversity in student performance levels to capture variability in feedback quality.

3.4 Evaluation Metrics and Scoring Model

A hybrid scoring model is developed, combining:

- Quantitative Scores (Likert scale-based evaluation across dimensions)
- Qualitative Thematic Analysis (identification of feedback patterns and depth)

Inter-rater reliability is assessed for faculty evaluations, while AI outputs are evaluated for consistency across multiple iterations, consistent with findings from Sreedhar et al. (2025).

3.5 Analytical Procedure

The analysis follows a multi-step process:

1. Content Normalization – Standardizing feedback length and format
2. Dimension-wise Scoring – Applying the analytical framework
3. Comparative Statistical Analysis – Identifying differences between AI and faculty feedback
4. Thematic Interpretation – Extracting qualitative insights

The methodology ensures both rigor and replicability, aligning with best practices in medical education research (Gordon et al., 2024).

4. RESULTS

The comparative analysis reveals distinct performance patterns between generative AI and faculty-derived feedback across the four evaluation dimensions.

In terms of linguistic quality, AI-generated feedback demonstrates superior performance, characterized by

high coherence, grammatical accuracy, and structured articulation. The consistency of language use across multiple responses indicates the robustness of generative models in producing standardized outputs. This aligns with previous findings suggesting that AI excels in producing well-structured educational content (Ali et al., 2026).

However, in clinical accuracy, faculty feedback outperforms AI-generated responses. Faculty members demonstrate a deeper understanding of psychiatric nuances, particularly in interpreting complex behavioral descriptions and contextual cues. AI-generated feedback, while generally accurate, occasionally exhibits superficial reasoning or overgeneralization, reflecting limitations in contextual comprehension.

The dimension of pedagogical relevance further highlights the strengths of faculty feedback. Educators provide targeted, actionable suggestions tailored to individual student performance, aligning with competency-based frameworks (Lee & Chiu, 2022). In contrast, AI-generated feedback tends to be generic, lacking personalization despite its structural clarity.

The most significant difference is observed in cognitive impact. Faculty feedback consistently promotes higher-order thinking by encouraging reflection, critical analysis, and clinical reasoning. AI feedback, although informative, often lacks the depth required to stimulate reflective learning processes. This finding reinforces concerns raised in the literature regarding the limitations of AI in fostering deep learning (Torous & Greenberg, 2025).

Quantitative analysis indicates that AI achieves higher scores in consistency and scalability, while faculty feedback scores higher in depth and educational effectiveness. Notably, the variance in faculty feedback is higher, reflecting differences in individual teaching styles and subjective interpretations.

These findings suggest that AI and faculty feedback systems possess complementary strengths. While AI offers efficiency and standardization, faculty feedback provides depth and contextual relevance. The results strongly support a hybrid feedback model integrating both approaches, consistent with emerging trends in AI-augmented education (Ali et al., 2026).

5. DISCUSSION

The findings of this study provide critical insights into the evolving role of generative artificial intelligence in medical education, particularly in the domain of narrative feedback for written Mental Status Examinations. The observed dichotomy between AI-generated and faculty-derived feedback underscores the fundamental differences in computational and human cognitive processes.

From a theoretical perspective, the superior linguistic performance of AI can be attributed to its training on large-

scale textual datasets, enabling it to generate syntactically accurate and coherent responses. This aligns with the conceptualization of AI as a tool for enhancing efficiency and standardization in educational processes (Boscardin et al., 2024). However, the limitations observed in clinical accuracy and cognitive impact highlight the challenges of applying AI in domains requiring deep contextual understanding and interpretive reasoning.

The findings corroborate the conclusions of Ali et al. (2026), which emphasize that AI-generated feedback, while structurally sound, lacks the nuanced judgment inherent in human educators. This limitation is particularly pronounced in psychiatry education, where interpretation of subjective experiences and behavioral cues is critical (Lee QY et al., 2025).

The pedagogical implications of these findings are significant. Effective feedback in medical education must not only convey information but also facilitate learning through reflection and critical thinking (Amonoo et al., 2021). The inability of AI to consistently achieve this suggests that it should not be viewed as a replacement for faculty but rather as a complementary tool.

The concept of a hybrid feedback model emerges as a key implication of this study. Such a model leverages the strengths of both AI and human educators, combining the scalability and consistency of AI with the depth and adaptability of faculty feedback. This approach aligns with the recommendations of Torous and Greenberg (2025), who advocate for augmenting rather than replacing traditional educational practices.

However, the integration of AI into feedback systems raises several challenges. Issues related to bias, transparency, and ethical use must be carefully addressed (Sallam, 2023). Additionally, the variability in faculty feedback highlights the need for standardized training and calibration to ensure consistency in educational outcomes.

Another critical consideration is the potential impact on student learning behavior. Over-reliance on AI-generated feedback may lead to passive learning, while excessive variability in faculty feedback may create confusion. Therefore, the design of feedback systems must balance consistency with personalization.

The study also identifies methodological limitations, including the restricted sample size and the focus on a single assessment type. Future research should explore longitudinal effects and extend the framework to other domains of medical education.

6. CONCLUSION

This study presents a comprehensive analytical

framework for evaluating generative AI and faculty-derived narrative feedback in medical students' written Mental Status Examinations. The findings demonstrate that while AI offers significant advantages in consistency, scalability, and linguistic quality, faculty feedback remains superior in clinical accuracy, pedagogical relevance, and cognitive impact.

The research contributes to the growing body of literature on AI in medical education by providing a structured approach to feedback evaluation. It also highlights the importance of integrating AI as a supportive tool rather than a replacement for human educators. The proposed hybrid model offers a practical pathway for optimizing feedback systems in medical training.

Future research should focus on refining the framework, exploring adaptive AI systems, and assessing long-term learning outcomes. The integration of AI into medical education must be guided by pedagogical principles, ethical considerations, and a commitment to enhancing student learning.

REFERENCES

1. 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
2. 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
3. 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
4. Çiçek FE, Ülker M, Özer M, Kyak 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
5. Duncan J. MSE Grader: GPT-4o. OpenAI; 2024. Accessed May 20, 2024.
6. 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
7. 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

8. 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
9. 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
10. 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
11. 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
12. 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.565448
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. 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
15. 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
16. 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
17. 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
18. 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.