Large language model used to simulate psychiatric OSCE scenarios: a medical student perspective

Zhaochu Geng,[1]* Craig S. Webster,[2,3] Yan Chen,[2] Lillian Ng,[4,5] Christian U. Krägeloh,[6] Angel Li,[1] and Marcus A. Henning.[2]

Affiliations:

[1]School of Medicine, University of Auckland, Auckland, New Zealand

[2]Centre for Medical and Health Sciences Education, School of Medicine, University of Auckland, Auckland, New Zealand

[3]Department of Anaesthesiology, School of Medicine, University of Auckland, Auckland, New Zealand

[4]Department of Psychological Medicine, Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand

[5]Auckland Regional Forensic Services, Health New Zealand, Te Whatu Ora, Auckland, New Zealand

[6]Department of Psychology, Auckland University of Technology, Auckland, New Zealand

*Correspondence to: zgen174@acuklanduni.ac.nz

Large language model used to simulate psychiatric OSCE scenarios: a medical student perspective

Abstract

AIM: This study investigates how varying prompt conditions influence the quality and clinical coherence of responses generated by a large language model (GPT-4o mini) in simulated psychiatric OSCE scenarios.

METHODS: Four psychiatric OSCE cases were presented to GPT-4o mini under four conditions with increasing details: a standard clinical prompt, a context-enhanced prompt, and two variation prompts incorporating irrelevant or distracting information. GPT-4o mini was asked to perform key OSCE tasks, history-taking, risk assessment, explanation, and management for each case. Responses were scored using a standardised, structured rubric and analysed thematically.

RESULTS: GPT-4o mini generated clinically relevant responses under standard and context-enhanced prompts. However, performance declined as irrelevant information was introduced. Quantitative scores dropped significantly across the different conditions, and qualitative analysis revealed reduced coherence, increased verbosity, and difficulty prioritising clinical content.

CONCLUSIONS: LLMs like GPT-4o mini can generate useful responses when provided with clear and concise prompt instructions. However, in this study, we noted that clinical accuracy and coherence deteriorated in the presence of distracting or ambiguous input. This highlights the need for critical evaluation and unambiguous literacy when using LLMs in medical education.

Keywords: Artificial intelligence; Large language models; Medical education; Objective structured clinical examination; Psychiatry education; Generative AI

Introduction

The rapid development of artificial intelligence (AI), particularly large language models (LLMs), is transforming multiple sectors, including healthcare and medical education.[1] The use of AI offers new opportunities for personalised learning, automated feedback, and simulated clinical experiences. Its potential to augment traditional teaching methods and support the training of future healthcare professionals is a growing area of academic and clinical interest.[2] This study directly evaluates the performance of a large language model (GPT-4o mini) in a high-stakes clinical exam, i.e., the Objective Structured Clinical Examination (OSCE). Understanding how LLMs perform in this context is crucial for their safe and effective integration into student learning.

OSCEs are a core component of medical education, designed to assess clinical skills in a standardised and objective manner.[3] In psychiatry, OSCEs are particularly valuable for evaluating history-taking, risk assessment, communication skills, and treatment planning in simulated patient encounters.[4] These assessments provide students with a structured environment to develop clinical reasoning and interpersonal skills through direct feedback.

Given the growing accessibility of LLMs, there is a compelling rationale to investigate whether these models can produce clinically accurate responses to medical scenarios. LLMs are increasingly used to simulate patient interactions and generate model answers, making them a potentially valuable resource for self-directed practice and formative exam preparation.

Despite its promise, the role of AI in practical clinical assessments such as OSCEs remains underexplored.[4] Soong and Ho suggest AI holds significant potential to reshape OSCE design

and delivery, although its application remains largely theoretical and untested. Further research is needed to assess its pedagogical, practical, and ethical implications.[5] This study seeks to address that gap by evaluating how a LLM performs across psychiatric OSCE scenarios.

As increasing numbers of medical students are turning to generative AI tools for OSCE preparation, it is important to understand not only whether these models can produce clinically relevant responses, but also how sensitive they are to variations in prompt structure and irrelevant detail. Evaluating this vulnerability is essential to determine whether LLM-generated outputs can be relied upon as consistent and safe supplementary learning tools in high-stakes assessment contexts.

This study aims to answer the following research question:

"How do varying prompt conditions affect the quality and clinical coherence of responses generated by GPT-4o mini in psychiatric OSCE scenarios?"

Materials and Methods

Study design and setting

This was a mixed-methods study conducted in a virtual setting to identify initial trends in GPT-4o mini performance of particular interest to medical students, and inform the scope and methodology for future, larger-scale research. The approach combined quantitative analysis of response quality with qualitative thematic analysis.

AI model and data collection

The LLM used in this study was GPT-4o mini (Open AI, San Francisco).[6] It was selected due to its accessibility and growing use in educational settings. To ensure consistency and minimise

the risk of variability from model updates, all GPT-4o mini responses were collected within a fixed seven-day period (5 January to 11 January 2025).

Four OSCE scenarios were selected from *oscesense.com*, an educational platform developed by qualified UK-based doctors.[7] The selected scenarios reflect common clinical presentations encountered by junior doctors and medical students in psychiatric training:

1. Scenario 1: A 23-year-old male presenting after a paracetamol overdose in a Liaison Psychiatry context.

2. Scenario 2: A 16-year-old girl presenting with mood and behaviour concerns in a General Outpatient Clinic.

3. Scenario 3: A 19-year-old girl presenting with weight loss and eating habits concerns in a GP clinic.

4. Scenario 4: A 20-year-old male presenting with complaints of poor sleep and difficulty concentrating in a GP clinic.

Each scenario required GPT-4o mini to perform standard OSCE tasks: taking a patient history, conducting a risk assessment, explaining the patient's risk to an examiner, and outlining a management plan.

During data collection, the first author maintained a personal memo documenting observations, reflections, and preliminary interpretations during each interaction. These notes were later reviewed alongside GPT-4o mini responses to support reflexive awareness and contextual understanding during thematic analysis.

Prompt conditions

To investigate the impact of contextual and distracting information on GPT-4o mini output, responses for each of the four scenarios were generated under four distinct prompt conditions (Table 1):

Table 1: Summary of prompt conditions used for psychiatric OSCE scenarios presented to GPT-4o mini.

Condition	**Description**
Original Condition	The standard, concise OSCE scenario typically encountered in examinations was presented.
Context Condition	The original scenario was presented with a brief instruction asking AI to adopt the role of a medical student sitting an OSCE. No additional environmental or personal details were provided.
Variation Condition 1	In addition to the Context Condition, this prompt included irrelevant environmental details (e.g., room temperature, ambient smells).
Variation Condition 2	In addition to Variation Condition 1, this prompt included irrelevant personal details about the patient (e.g., hobbies, fun facts) and explicit user-directed distractors (e.g., a glowing rubber duck on the examiner's desk).

This structured variation in conditions allowed for detailed analysis of how GPT-4o mini processes and prioritises information when presented with increasing levels of extraneous detail. Each prompt condition for each scenario was run three times to capture variability in AI outputs. Between the presentation of each individual prompt condition the investigator logged out of the session and logged back in again to clear the temporary cache of GPT-4o mini and eliminating the potential for order effects.

The quality of responses was assessed by the first author, a Year 3 medical student, who was mentored by a Year 5 medical student.

Data analysis

Quantitative analysis

LLM responses were evaluated and scored by the first author, using a predefined marking rubric developed by qualified UK-based doctors with expertise in medical education and psychiatry.[7] The rubric is aligned with standard OSCE assessment criteria and assessed the LLM responses across several domains. Each domain consisted of multiple objective criteria: points were awarded if the LLM explicitly responded to a criterion, and points were withheld if the criteria was not addressed. Domain scores were combined to produce a total score ranging from 0 to 100, allowing for an overall assessment of the LLM's clinical and communication performance.[7] Table 2 below presents the full marking rubric for Scenario 1.

Table 2: OSCE marking rubric for psychiatric scenario 1.

Domain	**Criteria**
Introduction	Adheres to appropriate infection control measures
	Introduces self with full name and role
	Correctly identifies patient using at least 2 patient identifiable variables
	Gains consent for history
History of overdose (Obtains clear history of overdose - including:)	Number of tablets taken
	Strength of tablets
	Taken at once or staggered

	When the last tablet was taken
	Establishes whether anything else was taken with the tablets
Risk assessment (Adequate assessment of risk to self - including:)	Protective factors
	Planning
	Location of attempt
	Patient's perceived lethality of attempt
	Patient's thoughts after attempt
	Whether patient was intoxicated
	Exploration of risk to others, pets or property
	Exploration of risk of others towards the patient
Risk categorisation	Establishes patient as high risk of harm to self, low risk of harm to pets and others - therefore overall high risk patient.
Assessment and management	Full psychiatric history
	Mental state exam
	Suggests review of the patient by a senior with possible view to detention of patient
	Suggests leasing with community mental health teams or crisis team
	Considers use of medication in the long term.
	Considers use of counselling services and talking therapy later down the line
	Candidate's management indicates that they are aware it would be unsafe to allow this patient to leave the hospital at present.

Note: Total score ranges from 0 to 100, with each criterion equally weighted at 4 points.

Descriptive statistics, including mean, median, standard deviation (SD), interquartile range (IQR), and 95% confidence intervals (CI), were calculated using Google Sheets across scenarios and prompt conditions.

Qualitative analysis

The first author adopted a qualitative approach, interpreting GPT-4o mini's responses simulating how a student would answer questions in an OSCE setting. This process reflected the author's personal observations and thought processes, which were documented informally as reflective notes taken during each interaction. Thematic analysis was conducted on the LLM generated responses following Braun and Clarke's established six-phase framework.[8]

Results

Overall AI performance

GPT-4o mini consistently generated responses that were well-structured and clinically relevant across all four psychiatric scenarios. However, a progressive deterioration in the overall response quality was observed as prompt complexity increased, with irrelevant information and explicit distractors affecting accuracy and completeness.

Quantitative findings

Quantitative analysis of the AI-generated responses provided a numerical evaluation of performance across the four prompt conditions. These data provide a direct comparison of how GPT-4o mini's response quality was affected by the introduction of irrelevant information.

Table 3: Summary statistics of GPT-4o mini performance scores by prompt condition, aggregated across four clinical scenarios (n = 4 per condition).

Condition	Mean score	Median score	Standard deviation	Interquartile range	95% confidence interval
Original	49	48	5.7	6.5	45.8 - 52.2
Context	53.4	53	5.1	9	50.5 - 56.3
Variation 1	43.9	44.5	5.9	8.5	40.6 - 47.2
Variation 2	40.4	42	4.9	8.5	37.6 - 43.2

Note: For each prompt condition, four independent clinical scenarios were analysed (n = 4 per condition). Three outputs were generated per scenario and averaged to produce a single score prior to quantitative analysis. Descriptive statistics were calculated across the four scenarios. Possible scores ranged from 0 to 100.

As shown in Table 3, both the mean and median scores progressively decreased from the Original and Context conditions to Variation 1 and Variation 2. The 95% confidence intervals (CIs) between the Original (45.8–52.2) and Variation 2 conditions (37.6–43.2), as well as between the Context condition (50.5–56.3) and both Variation conditions (40.6–47.2 and 37.6–43.2), show no overlap, suggesting a decline in performance across conditions.

In contrast, the standard deviation (SD) and interquartile range (IQR) remained relatively consistent across all four prompt conditions, indicating that the amount of variability in GPT-4o mini response quality did not significantly change. This suggests that complexity reduced the average quality of LLM outputs, but the degree of fluctuation in performance between individual responses remained stable.

Detailed descriptive statistics for each scenario and prompt condition can be viewed in Appendix A.

Qualitative findings

Our thematic analysis identified three key patterns in how GPT-4o mini response quality changed as the prompt complexity increased. These themes offer insights into the nuanced ways GPT-4o mini responses were affected by irrelevant information. Table 4 summarises the themes, their associated codes, and representative exemplar quotations from different prompt conditions (Original, Context, Variation 1, Variation 2) selected to illustrate each theme.

Table 4: Summary of qualitative themes, codes, and exemplars identified in LLM responses across prompt conditions.

Themes	**Codes**	**Exemplars from GPT-4o mini responses**			
		Original	**Context**	**Variation 1**	**Variation 2**
Complexity	Over-explaining simple tasks Redundancy in clinical explanations Reduced conciseness	*"Try to go to bed and wake up at the same time every day, even on weekends."* *"Avoid caffeine and alcohol, especially in the afternoon or evening."* *"Create a relaxing bedtime routine and avoid using electronic devices before*	*"Educate the patient on sleep hygiene techniques, such as:* *Regular sleep-wake times* *Avoiding screens 30 minutes before bed* *Avoiding caffeine and heavy meals late in the day* *Creating a relaxing bedtime*	*"Encourage the patient to establish a regular sleep schedule (go to bed and wake up at the same time every day).* *Suggest creating a relaxing bedtime routine (e.g., reading, avoiding electronics).* *Advise limiting*	*""One of the most effective ways to improve sleep and concentration is by making changes to your lifestyle. This could include setting a regular sleep routine, avoiding screens before bed, and creating a calm and relaxing environment at*

		bed."	*routine*."	*caffeine intake, especially in the afternoon. Recommend a quiet, dark, and cool sleep environment."*	*night. Limiting caffeine, alcohol, and heavy meals before bedtime can help as well. Physical activity can also improve both sleep quality and mood, so I'd recommend trying to get some exercise regularly.* "
Distraction	Acknowledgement of distractors Integration of non-clinical details Shift in tone	*"What's your typical day like? Do you feel supported by friends and family?"*	*"What do you typically do in a day, and how are you managing with school, work, or other responsibilities?"*	*"Engage her interest: "I heard you collect antique spoons. That's such a unique hobby—what's your favourite one so far?"*	*"I'm not sure if it's just me, but I've noticed that rubber duck over there... Is that some sort of magical artifact, or just a regular glowing toy? I apologize, I tend to get distracted sometimes."*
Lack of logic and	Abrupt topic transitions	*"What's been going on in your life recently?*	*"Can you think of any major stressors in your*	*"Are you still enjoying your hobbies, like*	"*You mentioned meeting Taylor Swift—how was*

connecti on	Disjointed sections Checklist-style progression	*Anything in particular that has been challenging for you?"*	*life recently, such as university, work, or personal relationships?"*	*collecting spoons? How about spending time with friends?"*	*that experience? It sounds fun! Has there been anything in your life lately that's been especially stressful or affecting your well-being?*"

Theme 1: Complexity

As the prompt complexity increased in the Variation conditions compared to the Original and Context conditions, GPT-4o mini's responses became more opaque, less structured and less concise. This was particularly evident in the management plans, which shifted from brief, directive statements to lengthy, list-style explanations. The exemplar from Variation 2 demonstrates an overly detailed and indirect delivery of lifestyle advice, lacking integration into a coherent, patient-focused explanation. While the content was clinically appropriate, the excessive elaboration and lack of conciseness in Variation Conditions 1 and 2 could be confusing in real-world clinical settings, where time is limited and communication must be clear. This suggests that the sheer volume and irrelevance of certain prompt details can subtly affect the directness and clarity of GPT-4o mini's output.

Theme 2: Distraction

GPT-4o mini's tone and clinical focus shifted as prompts included more irrelevant information. In the Original and Context conditions, the responses remained professionally and clinically oriented, focusing on understanding the patient's daily routine and social support. However, in

the Variation conditions, GPT-4o mini began incorporating non-clinical details such as the patient's spoon collection hobby into its responses, acknowledging and expanding on distractors in ways that disrupted the clinical flow. This shift in tone and attention can reduce the clarity, efficiency, and professionalism of the response, especially in time-pressured clinical scenarios where communication and judgement must remain succinct.

Theme 3: Lack of logic and connection

In the more complex Variation Conditions, GPT-4o mini responses occasionally exhibited a lack of logical flow and smooth transitions between different clinical questions. Instead of presenting a cohesive narrative, it sometimes jumped between unrelated topics or adopted a checklist-style structure that lacked integration. Distractors and irrelevant information may have contributed to the fragmentation of these responses. As shown by the exemplar in Variation Condition 2, there is a sudden shift from empathetic engagement to an important clinical question. This type of abrupt topic shift and lack of connection between clinical domains can make GPT-4o mini responses feel disjointed, reducing their overall coherence and potentially undermining the impression of professional competence.

Discussion

Interpretation of findings

This study provides relevant insight into the practical and clinical performance of a LLM (GPT-4o mini) when applied to defined psychiatric OSCE tasks under varying prompt conditions. Both quantitative and qualitative results suggest trends in which AI-generated responses tend to decrease in clinical quality as prompt complexity and irrelevant information increase.

Quantitatively, the mean and median scores decreased progressively from the Original and Context conditions to the more complicated Variation 1 and Variation 2 conditions. This

numerical trend supports the argument that although GPT-4o mini can generate structured, clinically meaningful responses, they are sensitive to "prompt noise", information that is irrelevant, distracting, or emotionally or contextually nuanced. These distractions might have impaired the GPT-4o mini's ability to prioritise key clinical elements, as seen in the drop in scores, consistent with Hadi et al. who found that ChatGPT's diagnostic performance can vary substantially depending on input phrasing and contextual detail.[9]

Qualitatively, the AI's increasing verbosity, lack of conciseness, and abrupt topic transitions appeared to reflect difficulty in distinguishing clinically relevant information from extraneous detail within the prompt. When it is faced with irrelevant or emotionally loaded information, such as a glowing rubber duck or the patient's spoon collection, GPT-4o mini seemed to struggle with filtering inputs and occasionally produced self-referential phrases, like a note to self or an apology for distraction. This behaviour likely reflects the model's probabilistic text-generation process, whereby it selects responses with the highest predicted likelihood based on its training data rather than through any form of cognitive effort or attentional control. As Geracitano et al. note, in general ChatGPT's apparent fluency often conceals substantial variability in accuracy and reasoning quality.[10] This could disrupt natural, fluent communication and undermine the professionalism necessary for high-stakes OSCE settings.

Overall, this performance pattern appears consistent with broader concerns about LLM's ability to navigate the inherently unpredictable nature of real-world clinical situations.[11,12] This reveals both the potential and limits of generative LLM in clinical education at this present time. It is excellent at linear, structured tasks but struggles with contextual noise, human nuance, and adaptive complexity.

Implications of findings

The findings of this study engender several important implications for how generative LLM, such as GPT-4o mini can be effectively and responsibly integrated into medical education, particularly in assisting students with their OSCE preparation.

1. Importance of prompt literacy for effective AI use

Our results demonstrate that the accuracy and clinical appropriateness of LLM-generated responses may be sensitive to the structure and relevance of input provided. This is consistent with Meskó, who highlights the need for students and educators to gain the skills of crafting clear and structured prompts to optimise LLM output.[13] Developing this form of prompt literacy will be important when interacting with LLM critically and precisely without compromised accuracy or safety, particularly in high-stakes educational contexts like OSCEs.

2. The role of AI in clinical training for medical students

GPT-4o mini appeared to be capable of generating structured responses involving history-taking, risk assessments, and basic management plans, particularly when provided with clean and well-structured prompts. AI's capacity to synthesise large volumes of clinical information makes it a potentially powerful resource for medical students, particularly in settings where access to mock exams, structured feedback, or simulated patients is limited.[14] Resources used for OSCE preparation such as Geeky Medics, OSCELab, and OSCER already utilise LLMs and natural language processing models to simulate clinical scenarios, offering a valuable supplement to traditional learning methods. These AI platforms can provide students with on-demand, low-cost learning opportunities, allowing them to enhance their clinical exposure and receive formative feedback outside of formal university settings.

However, the potential limitations of LLMs discussed above raise concerns about over-reliance on generative AI systems, particularly in under-resourced environments where medical

students may lack access to experienced educators for supervision or feedback.[15] Students should therefore be encouraged to critically appraise AI-generated content and remain aware of its limitations, especially in domains such as clinical judgement, empathy, and communication skills, which require human oversight. While AI holds considerable promise for enhancing accessibility and supporting medical education, it should be used as a supplement to, rather than a substitute for, human-led teaching.

3. Broader implications of AI in clinical practice and society

Beyond education, this study also touches on AI's potential role in clinical decision-making. As healthcare systems explore the use of LLM to support clinicians, particularly in high-pressure or resource-scarce environments, understanding the limitations of LLM reasoning becomes critical. The findings here echo broader concerns consistent with other studies indicating LLMs may appear confident while generating clinically irrelevant or biased outputs, especially when overwhelmed with contextually unimportant data.[16,17]

If AI is to assist healthcare professionals, it must first be refined to prioritise salient clinical information, reduce hallucinations, and operate with transparency and accountability. Without these safeguards, over-reliance on flawed outputs could compromise care quality and propagate existing disparities.

At the societal level, AI offers the promise of increased access to medical knowledge, especially for marginalised communities and individuals in remote or low-income settings. However, this requires parallel efforts to expand digital access, address AI-generated bias, and ensure regulatory oversight. Improved models, trained on diverse and high-quality clinical data, could help reduce disparities in health literacy, strengthen global medical education, and bring us closer to achieving health equity.[18]

Strengths and limitations

The strength of this study lies in its originality and its mixed method approach. It investigates a relatively underexplored area, LLM use in psychiatric OSCEs, and employs both quantitative scoring and thematic analysis to produce robust, multidimensional findings. It also draws from an authentic medical student lens that was moderated by medical education and clinical coauthors, adding relevance to how LLM might be integrated into actual student learning workflows.

However, the study has some limitations. First, all assessments were conducted by the primary researcher, without independent or blinded raters, which may introduce potential bias. In addition, although efforts were made to maintain consistent scoring criteria, the assessment process may have been influenced by the primary researcher's expectations regarding performance deterioration across prompt conditions. Nonetheless, the interpretation of the assessment material and processes were audited by medical education and clinical co-authors. Second, the lack of dynamic interaction between the LLM and simulated patients removes the ability to assess real-time conversational logic, empathy, or adaptability. Third, the sample size per condition (n=4) limits generalisability, and findings are model-specific to GPT-4o mini, which may differ from other LLMs in future performance.

Finally, GPT-4o mini was trained on a large and diverse dataset from multiple sources, but the exact composition, coverage of medical content, and recency of the data are not fully transparent. [19] This limits the ability to guarantee the accuracy of its medical outputs and suggests that observed performance may reflect both the model's training scope and inherent biases. Acknowledging these factors clarifies how the use of this generative AI model may influence the validity and generalisability of our findings.

Future research

Future research could aim to preserve the mixed-methods design of this study while enhancing methodological rigour and better simulating the complexities of real-time clinical practice. To reduce marker bias and improve generalisability, future studies would benefit from using larger sample sizes across diverse prompt conditions to improve statistical power and incorporate blinded, multi-rater assessments involving experienced clinicians as markers rather than medical students. Studies need to evaluate a broader range of LLM architectures, as comparisons across multiple models could clarify whether observed performance is model-specific or consistent across platforms. In addition, moving beyond static prompts to interactive OSCE simulations that allow real-time dialogue between LLMs, and simulated patients would enable assessment of conversational flow, empathy, and adaptive clinical reasoning.

Conclusion

This study offers insight into how GPT-4o mini performed under simulated OSCE conditions, highlighting both its promise and limitations in medical education. GPT-4o mini was able to generate clinically relevant and structured responses when given concise, well-formulated prompts, demonstrating its potential as a low-cost, accessible learning tool for students. However, its performance declined in the presence of irrelevant prompt information, raising concerns about its reliability in more complex, real-world scenarios. These findings underscore the importance of prompt literacy, critical engagement, and human oversight in the educational use of LLMs. While LLMs can enhance accessibility to clinical training, it should complement, not replace, human-led teaching, particularly for skills involving empathy, communication, and clinical judgement. As AI continues to evolve, further research is essential to ensure its integration into medical education remains evidence-based, ethically sound, and aligned with the core values of clinical training.

Conflicts of Interest: The authors declare no conflicts of interest.

Funding: This research was supported by a University of Auckland Summer Research Scholarship.

Disclosures: All authors had full access to the study data and take responsibility for the integrity and accuracy of the data analysis.

References

1. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. Med Sci Educ. 2025;35:1803-16. doi: 10.1007/s40670-025-02373-0

2. Nagi F, Salih R, Alzubaidi M, et al. Applications of artificial intelligence (AI) in medical education: a scoping review. Stud Health Technol Inform. 2023;305:648-651. doi: 10.3233/SHTI230581

3. Al-Hashimi K, Said UN, Khan TN. Formative objective structured clinical examinations (OSCEs) as an assessment tool in UK undergraduate medical education: a review of its utility. Cureus. 2023;15(5):e38519. doi: 10.7759/cureus.38519

4. Plakiotis C. Objective structured clinical examination (OSCE) in psychiatry education: a review of its role in competency-based assessment. Adv Exp Med Biol. 2017;988:159-80. doi: 10.1007/978-3-319-56246-9_13

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

6. OpenAI. GPT-4o mini [Internet]. San Francisco (CA): OpenAI; 2024 Jul [cited 2026 Jan 3]. Available from: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

7. Patel R, Pender S. Psychiatry OSCE stations [Internet]. OSCE Sense [cited 2025 Jul 6]. Available from: https://www.oscesense.com/psychiatry-osce-stations

8. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol. 2006;3(2):77-101. doi: 10.1191/1478088706qp063oa

9. Hadi A, Tran E, Nagarajan B, et al. Evaluation of ChatGPT as a diagnostic tool for medical learners and clinicians. PLoS One. 2024;19(7):e0307383. doi: 10.1371/journal.pone.0307383

10. Geracitano J, Anderson B, Coffel M, et al. The accuracy of ChatGPT in answering FAQs, making clinical recommendations, and categorizing patient symptoms: a literature review. Adv Health Inf Sci Pract. 2025;1(1):VXUL2925. doi: 10.63116/VXUL2925

11. Khan B, Fatima H, Qureshi A, et al. Drawbacks of artificial intelligence and their potential solutions in the healthcare sector. Biomed Mater Devices. 2023:1-8. doi: 10.1007/s44174-023-00063-2

12. Rao A, Pang M, Kim J, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. J Med Internet Res. 2023;25:e48659. doi: 10.2196/48659

13. Meskó B. Prompt engineering as an important emerging skill for medical professionals: tutorial. J Med Internet Res. 2023;25:e50638. doi: 10.2196/50638

14. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2(2):e0000198. doi: 10.1371/journal.pdig.0000198

15. Mohammad B, Supti T, Alzubaidi M, et al. The pros and cons of using ChatGPT in medical education: a scoping review. Stud Health Technol Inform. 2023;305:644-7. doi: 10.3233/SHTI230580

16. Griot M, Hemptinne C, Vanderdonckt J, et al. Large language models lack essential metacognition for reliable medical reasoning. Nat Commun. 2025;16(1):642. doi: 11.1038/s41467-024-55628-6

17. Roustan D, Bastardot F. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. 2025;14:e59823. doi: 10.2196/59823

18. Dankwa-Mullan I. Health equity and ethical considerations in using artificial intelligence in public health and medicine. Prev Chronic Dis. 2024;21:E64. doi: 10.5888/pcd21.240245

19. Zhui L, Fenghe L, Xuehu W, et al. Ethical considerations and fundamental principles of large language models in medical education: viewpoint. J Med Internet Res. 2024;26:e60083. doi: 10.2196/60083