Volume 4, Issue 3, March 2026 ISSN (E): 2938-3765

TEACHING CLINICAL REASONING TO MEDICAL STUDENTS USING AI-GENERATED CLINICAL SCENARIO TESTS: A MIXED-METHODS FORMATIVE EVALUATION

Mirzalieva Anora Arginbaevna
Teaching Assistant, Department of Propaedeutics of Internal Diseases No. 1.
Tashkent State Medical University, Tashkent, Uzbekistan

Kamalov Ruslan Kuralbaevich
3rd-Year Student
Tashkent State Medical University, Tashkent, Uzbekistan

Turgunboev Samandar Sanjar ugli
3rd-Year Student
Tashkent State Medical University, Tashkent, Uzbekistan

Norkulova Mubina Turgun kizi
3rd-Year Student
Tashkent State Medical University, Tashkent, Uzbekistan

Murodova Nigina Ilyos kizi
3rd-Year Student
Tashkent State Medical University, Tashkent, Uzbekistan

Abstract

The integration of artificial intelligence (AI) into medical education is rapidly evolving, offering new tools to enhance the teaching and assessment of clinical reasoning. One such tool is the Script Concordance Test (SCT), designed to evaluate clinical reasoning under conditions of uncertainty. Traditionally, developing SCT scoring systems and providing feedback requires the involvement of expert panels, which is time-consuming and resource-intensive. Recent advances in generative AI and large language models (LLMs) offer the potential to simulate expert judgment, yet this capability remains underexplored.

This study investigated the feasibility of using LLMs to emulate expert clinical judgment in the development, scoring, and feedback of SCTs in cardiology and pulmonology. Fifteen third-year medical students completed a 32-item test generated by ChatGPT-4o. Six LLMs were used as simulated experts, three of which were trained on course materials and three untrained. Students answered test items, rated perceived difficulty, and selected the most helpful feedback explanations. The average score was 22.8 out of 32. Trained models showed higher concordance with student

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

responses (ρ = 0.64) compared to untrained models (ρ = 0.41). AI-generated feedback was considered most useful in 62.5% of cases, particularly when using trained models.
These results indicate that trained generative AI models can reliably emulate expert clinical reasoning within the SCT framework. The use of such technology could simplify SCT development while maintaining educational value in feedback. Further research is needed to assess the long-term impact of these tools on the development of clinical reasoning and to determine the optimal balance between expert involvement and AI systems in medical education.

Keywords: Clinical reasoning; artificial intelligence; medical education; generative AI; simulation of expert judgment.

Introduction

One of the fundamental challenges in medical education is the assessment of clinical competencies, particularly higher-order skills such as clinical reasoning and decision-making. Tests that primarily focus on knowledge recall—such as multiple-choice questions—often fail to capture the full complexity of clinical reasoning under conditions of diagnostic uncertainty [1]. To address this, various assessment methods have been developed, including workplace-based assessments and the Objective Structured Clinical Examination (OSCE), both of which simulate real clinical interactions. However, these approaches are resource-intensive, requiring significant organizational effort and faculty involvement. An additional tool for assessing reasoning in uncertain clinical situations is the Script Concordance Test (SCT). Unlike traditional assessment formats, the SCT requires students to use a Likert-type scale to evaluate how new information presented within a clinical scenario impacts a proposed hypothesis. Scoring is based on the degree of concordance between students' responses and those of a reference panel of experienced clinicians, thereby reflecting the variability and probabilistic nature of real-world clinical practice [2,3]. Despite their educational value, developing and scoring SCTs is time-consuming and typically requires the involvement of multiple experts who must provide carefully considered and consistent responses to a large number of items.

A typical SCT format presents a clinical case followed by a diagnostic, investigative, or therapeutic hypothesis. New clinical information is then introduced, and the examinee must use a Likert-type scale to judge how this information affects the likelihood or relevance of the proposed hypothesis. Responses are scored by comparing the examinee's choices with those of a panel of experienced clinicians; partial credit is awarded based on the distribution of expert responses, rather than on the existence of a single correct answer. This scoring method accounts for the inherent variability in expert judgment and makes it possible to measure how closely students' reasoning aligns with expert clinical thinking, rather than simply testing factual knowledge. Evidence supporting the validity of SCTs includes their ability to distinguish between different levels of training, their positive impact on learners' cognitive engagement, and their acceptable reliability when a sufficient number of items and expert panel members are used [3–5]. Currently, SCTs are widely used in both undergraduate and postgraduate medical education across many countries, particularly in specialties requiring complex clinical decision-making.

In recent years, advances in artificial intelligence (AI)—especially the development of large language models (LLMs)—have begun to transform this landscape. LLMs are deep learning models trained on vast textual datasets; they can generate human-like text, synthesize complex concepts, and even simulate specialized professional reasoning [6]. Unlike traditional rule-based systems, these models produce context-dependent responses with high linguistic coherence, identifying patterns in unstructured data. Research has demonstrated that LLMs can simulate patient–doctor dialogues, generate high-quality educational feedback, and achieve performance at or near the passing threshold on tasks analogous to the United States Medical Licensing Examination (USMLE) [7–9].

In the context of using LLMs to generate SCTs, several studies have already highlighted the utility of this tool for developing clinical scenarios in medical education and other healthcare fields [10,11]. A recent study found that scenarios for SCTs generated using ChatGPT (OpenAI) were of high educational quality [11]. For generative AI tools to be effectively integrated into the development of educational materials—and to optimize the use of human and material resources—it is essential to assess the difficulty level of the scenarios and questions they produce [12]. Furthermore, to the best of our knowledge, no studies to date have demonstrated the ability of generative AI to function as a clinical expert in selecting the most appropriate response to a given scenario and providing feedback to learners.

The primary aim of this project is to investigate the effectiveness of generative AI in creating script concordance tests for medical students and in simulating the role of an expert. A secondary objective is to explore how medical students perceive the difficulty level of AI-generated SCTs and the usefulness of different types of feedback provided.

Methods

Study Design and Setting. This study employed a cross-sectional, mixed-methods experimental design to evaluate medical students' performance on a formative Script Concordance Test (SCT) developed using generative artificial intelligence, and to explore the feasibility of using AI systems as content experts for both scoring and providing formative feedback.

Participants and Recruitment. The study was conducted in February 2026 at the Department of Propaedeutics of Internal Diseases No. 1, Tashkent State Medical University. Eligibility criteria required participants to be third-year medical students who had completed the relevant coursework. This timing was chosen deliberately to allow for the inclusion of clinical reasoning questions addressing symptoms and signs related to both the cardiovascular and respiratory systems, given their close physiological and clinical interconnections.

Test Development and Structure. The Script Concordance Test was developed using generative AI (ChatGPT-4o) in accordance with established guidelines for SCT construction [13]. Four clinical scenarios were generated: acute exacerbation of chronic obstructive pulmonary disease (COPD), hemoptysis, chest pain, and acute cough. Each scenario contained eight items, resulting in a total of 32 questions. To ensure a consistent structure across scenarios, standardized prompts were used, specifying a 5-point Likert scale (ranging from "very unlikely" to "very likely" to reflect the impact

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

of new information on the hypothesis) and requiring logical coherence within the cardiopulmonary context.

Each item followed the classic SCT structure:

1. A clinical scenario containing an element of uncertainty;
2. A diagnostic, investigative, or therapeutic hypothesis;
3. A new clinical finding or piece of information.

Items were designed and reviewed to reflect varying levels of diagnostic uncertainty, probabilistic clinical reasoning, and realistic clinical situations.

The following prompts were used during SCT generation:

- Assume the role of a university-level health sciences education expert with specialization in cardiology and pulmonology;
- Assume the role of a specialist in developing Script Concordance Tests;
- Generate an SCT clinical scenario containing 8 questions on a topic integrating pulmonology and cardiology;
- Frame the questions as follows: "If you are considering '[diagnostic hypothesis]' and you discover '[symptom or sign]', how does this finding affect the likelihood of this hypothesis?";
- Use a Likert scale from -2 to +2 (from "very unlikely" to "very likely").

Although the AI-generated items were designed according to the classic SCT format (clinical scenario, hypothesis, new information, and Likert-scale rating), some questions did not fully adhere to this structure. This occurred when the generative model incorporated redundant information from the scenario into subsequent questions, occasionally blurring the boundary between the initial context and the new clinical information. Since the prompt did not explicitly restrict the reuse of previously presented information, this likely led to partial deviations from the intended format.

Using Generative AI as an Expert Panel. The reference expert panel consisted of six large language models (LLMs), selected to ensure diversity in reasoning patterns while maintaining practical feasibility. Although the classic SCT format typically involves 15–20 expert clinicians to achieve sufficient response variability [5], LLMs can rapidly generate consistent probabilistic estimates across multiple iterations. Pilot testing indicated that using six distinct model architectures provided adequate reasoning diversity for constructing an aggregated scoring key while keeping the process computationally manageable. The models were divided into two groups:

Trained AI experts (n = 3). These included ChatGPT-4o, Claude 3.7, and Microsoft 365 Copilot. They did not undergo additional fine-tuning at the parameter level. Instead, each model received contextual information in the form of course materials (lecture notes and clinical guidelines from the cardiopulmonary block), which were included directly in the prompt context. This approach—known as in-context learning or prompt-based domain adaptation—allowed the models to draw on relevant instructional information without modifying their underlying architecture.

Untrained AI experts (n = 3). The same base models were used, but without the addition of course materials or contextual information, reflecting their standard, general-purpose configuration.

Each AI system was required to answer all 32 SCT items, acting as an experienced clinical expert. The responses obtained were used to construct scoring keys and to generate feedback explanations

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

for each question. The modal response across all six AI experts was used as the reference answer for scoring purposes.

The prompts for generating feedback were identical across all AI systems and included the following instructions:

- Assume the role of a medical expert in cardiology and pulmonology;
- Assume the role of an SCT expert;
- For the given scenario and associated questions, select the most appropriate Likert-scale response;
- For each response, provide an explanation as if you were explaining the reasoning to a medical student.

Student Test Administration and Data Collection. Participants completed the 32-item AI-generated SCT via a Google Forms survey. For each item, students were asked to:

- Select a response on the Likert scale;
- Rate the perceived difficulty of the question on a 7-point scale;
- Choose the most and least helpful feedback explanation from among six anonymized AI-generated comments;
- Indicate whether they believed the question was created by a human or by AI.

Open-ended comment fields were also provided to collect qualitative feedback. Participation was voluntary and anonymous, with consent implied upon form submission. The estimated completion time was 35–45 minutes. Responses were exported and prepared for analysis in Microsoft Excel.

Scoring and Statistical Analysis. Student responses were scored using an aggregated partial-credit method:

- Full credit (1 point) was awarded if the student's response matched the modal expert response.
- Partial credit (e.g., 0.75 or 0.5) was awarded if the response matched a minority choice among the AI experts.
- Zero points were awarded if the response did not match any of the options selected by the expert panel.

Descriptive statistics—including overall scores, item-level performance, and perceived difficulty—were calculated. Agreement among AI systems, as well as between AI and student responses, was assessed using Spearman's rank correlation (ρ).

For the feedback analysis, student preferences regarding expert explanations were aggregated across all items. For each "expert" (AI system), a frequency score was calculated indicating how often its explanation was selected as most helpful or least helpful, allowing for a comparison between trained and untrained AI models as sources of feedback.

Qualitative Analysis

Open-ended responses were analyzed using an inductive thematic approach. Comments were reviewed sequentially, coded, and grouped into main and subthemes by two independent researchers. Discrepancies were discussed until consensus was reached. Themes addressed students' perceptions of the SCT format, the quality of the feedback provided, and suggestions for test improvement.

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

Results

Participants. A total of 25 students voluntarily participated in the study, completing the Script Concordance Test (SCT) under formative, non-graded conditions. For each item, students rated the perceived difficulty, selected the most and least helpful expert feedback explanations, and provided qualitative comments on individual questions as well as on the test as a whole. Six generative AI systems (three trained on course materials and three untrained) also completed the test and provided item-by-item rationales, which were subsequently used as immediate feedback for students.

Student Performance. Among the 25 third-year medical students who completed the SCT, the mean score was 22.8 out of 32, with individual scores ranging from 19.75 to 26.75. The score distribution approximated a normal curve with slight leftward skewness, indicating a reasonably high level of concordance between student responses and the AI expert panel, while still reflecting some variability in performance. This range of scores suggests that the AI-generated SCT was capable of differentiating levels of clinical reasoning among participants, confirming its discriminatory power even in a formative testing context.

Table 1. Student Performance on AI-Generated Script Concordance Test Items (N = 25)

Item	Item Text	Mean (SD)	Range (Min–Max)	Median (IQR)
1	How does an oxygen saturation of 85% affect the likelihood of an acute exacerbation of COPD?	0.72 (0.15)	0.25–1	0.75 (0.75–0.75)
2	How does a recent respiratory infection influence the likelihood of acute pulmonary edema?	0.65 (0.20)	0–0.75	0.75 (0.75–0.75)
3	How does the presence of dyspnea and tachycardia affect the likelihood of pulmonary embolism?	0.76 (0.17)	0.25–1	0.75 (0.75–0.75)
4	To what extent does a high heart rate influence the likelihood of myocardial dysfunction?	0.56 (0.27)	0–1	0.75 (0.50–0.75)
5	How important is chest X-ray for differentiating between pulmonary edema and an acute asthma exacerbation?	0.89 (0.21)	0.5–1	1 (1–1)
6	How does a history of smoking affect the likelihood of an acute COPD exacerbation?	0.82 (0.15)	0.5–1	0.75 (0.75–1)
7	How important is taking a patient history in the diagnosis of respiratory failure?	0.97 (0.11)	0.5–1	1 (1–1)
8	How do signs of acute ischemia on ECG influence the diagnosis of acute pulmonary edema?	0.79 (0.19)	0.25–1	0.75 (0.75–1)

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

Item	Item Text	Mean (SD)	Range (Min–Max)	Median (IQR)
9	How do night sweats and weight loss affect the likelihood of tuberculosis in the differential diagnosis?	0.78 (0.08)	0.75–1	0.75 (0.75–0.75)
10	How important is a history of bronchiectasis in assessing the risk of chronic hemoptysis?	0.76 (0.13)	0.25–1	0.75 (0.75–0.75)
11	How does an oxygen saturation of 92% influence the assessment of severe, life-threatening hemoptysis?	0.59 (0.32)	0–1	0.75 (0.25–0.75)
12	How does a 30 pack-year smoking history affect the likelihood of bronchogenic carcinoma?	0.86 (0.13)	0.75–1	0.75 (0.75–1)
13	How important is history-taking in identifying an infectious cause (e.g., tuberculosis or pneumonia)?	1.00 (0)	1–1	1 (1–1)
14	How important is chest X-ray in clarifying the cause of hemoptysis?	1.00 (0)	1–1	1 (1–1)
15	How does the interaction between smoking history and complication risk influence bronchiectasis activity?	0.82 (0.14)	0.5–1	0.75 (0.75–1)
16	How does a history of tuberculosis contact affect the likelihood of tuberculous hemoptysis?	0.86 (0.13)	0.75–1	0.75 (0.75–1)
17	How does ST-segment elevation in leads II, III, and aVF affect the likelihood of acute coronary syndrome?	0.99 (0.05)	0.75–1	1 (1–1)
18	How does diabetes mellitus influence the presentation of atypical acute coronary syndrome?	0.75 (0.16)	0.5–1	0.75 (0.75–0.75)
19	To what extent does arterial hypertension affect the likelihood of aortic dissection in a patient with chest pain?	0.77 (0.10)	0.5–1	0.75 (0.75–0.75)
20	How does pain radiation influence the likelihood of pneumothorax?	0.29 (0.27)	0–1	0.25 (0.25–0.25)
21	How important is ECG data in differentiating between acute coronary syndrome and pulmonary embolism?	0.96 (0.12)	0.5–1	1 (1–1)
22	How does elevated blood pressure affect the likelihood of pulmonary embolism?	0.55 (0.23)	0.25–0.75	0.75 (0.25–0.75)

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

Item	Item Text	Mean (SD)	Range (Min–Max)	Median (IQR)
23	How do sex-based differences in pain radiation influence the assessment of acute coronary syndrome likelihood?	0.77 (0.20)	0.25–1	0.75 (0.75–1)
24	How do diabetes mellitus and arterial hypertension influence the urgency of care?	0.78 (0.21)	0.5–1	0.75 (0.50–1)
25	How does the presence of initial fever affect the likelihood of community-acquired pneumonia?	0.86 (0.13)	0.75–1	0.75 (0.75–1)
26	How does ACE inhibitor use influence the development of treatment-related cough?	0.72 (0.18)	0.25–1	0.75 (0.75–0.75)
27	How does a normal oxygen saturation of 97% affect the assessment of severe pneumonia requiring hospitalization?	0.15 (0.16)	0–0.5	0.25 (0–0.25)
28	How does the presence of wheezing influence the diagnosis of acute bronchitis?	0.43 (0.28)	0–1	0.25 (0.25–0.75)
29	How important is history-taking in differentiating between ACE inhibitor-induced cough and acute respiratory infection?	1.00 (0)	1–1	1 (1–1)
30	How does a cough duration of 10 days affect the likelihood of an upper respiratory tract infection?	0.70 (0.20)	0.25–1	0.75 (0.75–0.75)
31	How does chest X-ray help differentiate between pneumonia and bronchitis?	1.00 (0)	1–1	1 (1–1)
32	How important are initial fever and productive cough in deciding to prescribe antibiotics for suspected pneumonia?	0.84 (0.20)	0.25–1	1 (0.75–1)

Abbreviations: SD, standard deviation; IQR, interquartile range; COPD, chronic obstructive pulmonary disease; ECG, electrocardiogram; aVF, augmented vector foot (lead aVF); ACE, angiotensin-converting enzyme.

Item-Level Analysis. A more detailed examination of item-level performance revealed distinct patterns across different types of clinical questions. Items featuring straightforward clinical findings—such as Q1 (assessing the impact of hypoxemia [oxygen saturation 85%] on the hypothesis of COPD exacerbation) and Q8 (evaluating the influence of ECG findings on the diagnosis of acute pulmonary edema)—demonstrated high concordance with the expert panel. For these items, over 80%

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

of student responses aligned with the modal expert choice, indicating confident application of established diagnostic patterns.

In contrast, items such as Q2 (the influence of recent respiratory infection on the hypothesis of acute pulmonary edema) and Q4 (the significance of tachycardia in suspected myocardial dysfunction) showed greater response variability, with fewer than 50% of students matching the most frequent expert response. These items involved a higher degree of clinical uncertainty and required more complex inferential reasoning, accounting for the wider distribution of responses.

Approximately 40% of students achieved scores of 25 or higher, demonstrating concordance with expert responses even on more ambiguous items. These students typically rated items as less difficult and made more judicious use of the "slightly confirmed" category on the Likert scale. The remaining 60% of students scored between 19 and 24 and more frequently selected "non-influenced" or "slightly confirmed" options, reflecting either a more cautious reasoning style or uncertainty in applying new clinical information. Overall, the SCT proved valuable in revealing variations in clinical reasoning styles and diagnostic confidence levels among novice clinical students.

AI-Student Concordance. To assess the extent to which AI models could simulate expert reasoning, we examined concordance—the degree of alignment between AI and student response patterns—using Spearman's rank correlation coefficient (ρ). This nonparametric measure evaluates the strength and direction of association between two ranked datasets: values approaching +1 indicate strong agreement, while values near 0 suggest weak or no association.

Among the trained AI systems, ChatGPT-4o showed the highest concordance with students ($\rho = 0.68$; $P < .001$), followed by Claude ($\rho = 0.64$) and Microsoft Copilot ($\rho = 0.61$). These values reflect moderately strong positive correlations, indicating that the trained models reasoned in ways consistent with typical student decision-making. In contrast, untrained models demonstrated weaker concordance (mean $\rho = 0.41$), reflecting less consistent or more generic reasoning patterns.

From a pedagogical perspective, these findings suggest that contextualized AI models—those provided with course materials via prompting—can effectively mirror how students weigh diagnostic information. This alignment supports the potential use of trained AI not only as an assessment tool but also as a feedback mechanism capable of modeling clinically sound reasoning.

Untrained models exhibited lower concordance levels. The untrained version of ChatGPT-4o showed moderate correlation ($\rho = 0.48$; $P = .04$), while untrained Claude and Copilot demonstrated weaker, statistically non-significant correlations (0.42; $P = .06$ and 0.34; $P = .08$, respectively). These models displayed less consistent patterns, often selecting neutral or non-committal responses ("non-influenced"), particularly on items requiring nuanced interpretation of clinical findings. Their reasoning appeared less contextually grounded and showed greater variability across similar clinical scenarios.

Perceived Item Difficulty. In addition to completing the 32 SCT items, students rated the perceived difficulty of each question on a 7-point Likert scale (1 = "very easy," 7 = "very difficult"). The mean difficulty rating across all items was 3.7, indicating moderate overall test difficulty appropriate for a formative assessment—challenging enough to stimulate clinical reasoning while remaining accessible.

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

At the item level, Q2 (influence of recent respiratory infection on the hypothesis of acute pulmonary edema) and Q4 (significance of tachycardia in suspected myocardial dysfunction) emerged as the most difficult, with mean ratings exceeding 4.5 (SD = 0.9). These items involved indirect or ambiguous relationships between new information and the clinical hypothesis, requiring reasoning within the "gray zone" of clinical data. Conversely, items with more straightforward diagnostic anchors—such as Q1 (hypoxemia in COPD) and Q8 (ECG ischemia in pulmonary edema)—were rated substantially easier, with mean difficulty below 3.0. This pattern confirms that item clarity and familiarity with pathophysiological mechanisms strongly influence perceived difficulty.

No statistically significant differences were found in perceived difficulty between items that students believed to be human-generated versus AI-generated (t-test; P = .47). All items were in fact AI-generated, indicating that students perceived the content as authentic. In open-ended comments, several students noted that they could not distinguish AI-generated questions, further supporting the high realism of the content generation.

Cluster analysis of difficulty ratings revealed differences between student subgroups. High-performing students (SCT score ≥25/32) rated the test as less difficult overall (mean 3.3, SD = 0.8) compared to their peers (mean 3.9, SD = 0.8), potentially reflecting greater familiarity with clinical reasoning patterns and more efficient script activation. In their comments, students emphasized that the primary challenge lay not in question length or wording, but in "determining the most relevant data" when faced with partially conflicting clinical findings.

Spearman correlation analysis between difficulty ratings and concordance with trained AI showed a moderate positive trend ($\rho = 0.32$; P = .08), suggesting that items more closely aligned with trained AI reasoning were perceived as less difficult. This finding supports the notion that AI concordance with instructional scenarios facilitates cognitive processing for students.

Evaluation of AI-Generated Feedback. For each question, students selected the most and least helpful explanations from among six AI-generated options (trained and untrained models). Their selections were guided by three criteria: clarity, clinical relevance, and educational value.

Across all 32 items, AI-generated feedback was rated as helpful in 62.5% of cases, particularly when generated by models trained on course materials. Trained ChatGPT-4o was most frequently selected, followed by trained Claude and trained Copilot. Students cited clarity, conciseness, and alignment with taught clinical reasoning strategies as reasons for their preferences. Feedback from trained models was especially valued on items requiring pathophysiological reasoning, such as those involving acute coronary syndromes and differential diagnosis of respiratory conditions. The consistent logic and evidence-based approach enhanced the credibility and educational value of the explanations.

In contrast, feedback from untrained models was rated as most helpful in only 9.4% of cases and was more frequently selected as least helpful. Untrained models often provided generic or overly cautious explanations, using phrases such as "may support the hypothesis" or "further investigation is required," without accounting for the specific clinical data presented in the scenario. At times, they failed to recognize key pathophysiological relationships, reducing the relevance of their feedback to students' learning needs. While not outright incorrect, the untrained AI responses lacked the pedagogical precision necessary for effective formative feedback.

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

Discussion

Main Findings. This study investigated the ability of generative AI to create Script Concordance Tests (SCTs) for medical students and to serve as an expert entity—fully developing the test, scoring responses, and providing feedback. Among the 25 third-year medical students who completed the AI-generated SCT, the mean score was 22.8 out of 32, with the distribution of results reflecting meaningful differences in clinical reasoning patterns. Students generally showed good concordance with the expert-modeled responses, particularly on items with clear pathophysiological anchors, while questions requiring more complex inferential reasoning elicited greater response variability. Trained AI models (those provided with course materials) demonstrated the highest concordance with both student performance and expected reasoning pathways, delivering feedback that was most frequently rated as helpful. In contrast, untrained AI models were perceived as less pedagogically effective and showed weaker correlations with student responses. The SCT also demonstrated acceptable reliability (Cronbach's $\alpha = 0.76$) and was rated by students as moderately difficult, further supporting its ability to capture meaningful differences in clinical reasoning. Interestingly, preliminary trends suggested that students with higher SCT scores more frequently selected feedback from trained AI models, indicating that the pedagogical structure of these explanations may resonate better with learners who already demonstrate stronger reasoning skills. This observation supports the notion that trained AI could help reinforce expert reasoning scripts, although larger-scale studies are needed to confirm these findings.

Comparison with Previous Studies. The integration of AI into formative assessment tools for clinical skills—such as SCTs—is still in its early stages, but this study provides a practical demonstration of how LLMs can model elements of expert reasoning within the SCT framework when properly contextualized. While previous work has already demonstrated the clinical reasoning capabilities of LLMs under various testing conditions (e.g., Nori et al. [14], Singhal et al. [6]), the present study extends these findings into formative education by integrating AI-generated questions, scoring, and feedback into a unified workflow. Whereas prior research has focused on LLMs' ability to answer multiple-choice questions or USMLE-style exams—formats that primarily rely on knowledge recall and fixed responses [15,16]—SCTs require probabilistic reasoning and tolerance for uncertainty, qualities that more closely mirror real-world clinical thinking and can be used to assess clinical skills. The ability of trained LLMs in this study to replicate the distribution of expert responses and provide comprehensible, pedagogically meaningful explanations suggests that AI may serve not only as a content generator but also as a cognitive analogue to an experienced clinician.

Furthermore, these findings align with recent research demonstrating AI's capacity for reflective (metacognitive) reasoning. For instance, Nori et al. [14] noted that LLMs such as GPT-4 are increasingly capable of adapting their rationales based on context and task complexity. Our data support this, particularly on items where trained models provided nuanced, detailed explanations that resonated with learners. Moreover, the fact that students could not distinguish AI-generated feedback from expert feedback is consistent with the work of Lee and Song [17], who showed that students often fail to differentiate between AI and expert explanations when content quality is high. Importantly, students demonstrated a willingness to accept AI-generated feedback provided it was

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

contextually accurate and well-structured, indicating the potential for trusted AI integration in educational settings.

At the same time, this project illustrates the current limitations of general-purpose LLMs. Untrained models, while producing grammatically correct responses, often delivered cautious or generic explanations, particularly on items requiring nuanced interpretation of subtle clinical findings. This diminished their pedagogical value and student trust. These observations echo the critique by Arora and Arora [18], who argue that untrained LLMs lack the necessary contextual grounding for expert-level interpretation, especially in nuanced or culturally specific clinical situations. Thus, effective AI integration into medical education requires model customization, thoughtful prompt engineering, and ongoing human oversight to ensure accuracy, relevance, and educational value of explanations.

Limitations

This study has several limitations. The sample size was small and confined to a single institution, which may limit the generalizability of the findings. The test was formative in nature, and students may not have invested the same level of cognitive effort as they would in a summative assessment. It is also possible that participation was skewed toward students with an interest in AI, potentially influencing the qualitative results. Additionally, while the AI models were trained on course materials, their responses were dependent on prompt accuracy and token limitations, which may have affected outcomes. A direct comparison with a panel of human experts on the same test was not conducted, limiting our ability to assess whether AI–student concordance reflects true expert clinical judgment. Consequently, the results demonstrate internal validity and feasibility but not equivalence to expert evaluation. Future studies should include parallel expert panels to establish criterion validity and calibrate the pedagogical quality of AI-generated scoring and feedback. As this was a single-administration study, the long-term impact on clinical reasoning development and knowledge retention was not assessed. Future research should examine how repeated exposure to AI-generated SCTs influences learning trajectories and whether hybrid models (AI + human) outperform AI alone. Finally, the risk of AI "hallucinations" or inaccurate information in generated explanations must be considered. Although plausibility and internal consistency were checked, formal content validation against reference sources was not performed. As Masters [19] notes, generative AI systems can produce fabricated or misleading medical information—a phenomenon known as "AI hallucination." This risk underscores the need to treat AI-generated explanations as adjunctive rather than authoritative, particularly in formative learning contexts. Future implementations should include systematic expert review to ensure clinical accuracy and safeguard students against misinformation.

Conclusions

As AI technologies continue to evolve and become increasingly integrated into medical education, this study demonstrates new possibilities for leveraging large language models (LLMs) in roles traditionally occupied by human experts in the assessment of clinical reasoning. By developing and scoring an SCT entirely through generative AI models, and by employing both trained and untrained models to simulate expert feedback, we have shown that AI can generate assessment materials that are functionally valid and positively received by students. The high degree of concordance between student responses and trained AI panels, together with favorable student evaluations of AI-generated

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

explanations, points to the genuine potential of these technologies to support formative assessment. Despite the performance gap between trained and untrained models, the study indicates that, with appropriate customization, AI can enhance both the efficiency and pedagogical value of clinical training.

These findings open the door to further research into scalable, AI-supported assessment models capable of flexibly supporting reasoning skills under conditions of uncertainty—a core competency for the contemporary physician. Beyond demonstrating feasibility, this work highlights the transformative potential of generative AI to democratize access to high-quality formative assessment. In academic environments where resources and expert availability are limited, leveraging AI for SCT generation could substantially reduce faculty workload, expand curriculum coverage, and enable rapid adaptation of assessments to evolving learning objectives. By lowering the resource barriers traditionally associated with expert-driven evaluation, this approach positions AI as a potential catalyst for sustainable and scalable innovation in medical education on a global scale.

References

1. Schuwirth LWT, Van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. Med Teach. 2011;33(6):478-485. [CrossRef] [Medline]
2. Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The Script Concordance Test: a tool to assess the reflective clinician. Teach Learn Med. 2000;12(4):189-195. [CrossRef] [Medline]
3. Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script Concordance Test: a review of published validation evidence. Med Educ. 2011;45(4):329-338. [CrossRef] [Medline]
4. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Med Educ. 2012;46(6):552-563. [CrossRef] [Medline]
5. Gagnon R, Charlin B, Coletti M, Sauvé E, van der Vleuten C. Assessment in the context of uncertainty: how many members are needed on the panel of reference of a script concordance test? Med Educ. 2005;39(3):284-291. [CrossRef] [Medline]
6. Singhal K, Azizi S, Tu T, et al. Publisher correction: large language models encode clinical knowledge. Nature. 2023;620(7973):E19-E19. [CrossRef]
7. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817-2825. [CrossRef] [Medline]
8. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al. Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023;9:e48291. [CrossRef] [Medline]
9. 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. [CrossRef] [Medline]
10. Kıyak YS, Emekli E. Using large language models to generate script concordance tests in medical education: ChatGPT and Claude. Rev Esp Edu Med. 2024;6(1). [CrossRef]

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5

11. Hudon A, Kiepura B, Pelletier M, Phan V. Using ChatGPT in psychiatry to develop script concordance tests in undergraduate medical education: a mixed methods study. JMIR Med Educ. 2024;10:e54067. [CrossRef] [Medline]
12. Choi GW, Kim SH, Lee D, Moon J. Utilizing generative AI for instructional design: a SWOT analysis. TechTrends. 2024;68(4):832-844. [CrossRef]
13. Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. BMC Med Inform Decis Mak. 2008;8:18. [CrossRef] [Medline]
14. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv. Preprint posted online April 12, 2023. [CrossRef]
15. Luo D, Liu M, Yu R, et al. Assessing the performance of GPT-3.5, GPT-4, and GPT-4o on the Chinese National Medical Licensing Examination. Sci Rep. 2025;15(1):14119. [CrossRef]
16. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. JMIR Med Educ. 2023;9:e45312. [CrossRef] [Medline]
17. Lee S, Song KS. Faculty and student perceptions of AI-generated explanations: insights for integrating generative AI in computer science education. Comput Educ Artif Intell. 2024;7:100283. [CrossRef]
18. Arora A, Arora A. The promise of large language models in health care. Lancet. 2023;401(10377):641. [CrossRef]
19. Masters K. Medical Teacher's first ChatGPT's "hallucinations": lessons for editors, reviewers, and educators. Med Teach. 2023;45(7):673-675. [CrossRef]

Web of Medicine: Journal of Medicine, Practice and Nursing
webofjournals.com/index.php/5