

Accuracy and Limitations of AI Scoring for Objective Medical Interview Checklist Items in a Year-One School of Medicine Objective Structured Clinical Examination (OSCE)

Richard L. Briley¹

r1briley@utmb.edu

The University of Texas Medical Branch at Galveston

Kelly R. Grace¹

The University of Texas Medical Branch at Galveston

Bruce W. Adcock¹

The University of Texas Medical Branch at Galveston

Erin C. King¹

The University of Texas Medical Branch at Galveston

Amanda D. Ripple¹

The University of Texas Medical Branch at Galveston

Erin K. McGoff¹

The University of Texas Medical Branch at Galveston

Patricia D. Self¹

The University of Texas Medical Branch at Galveston

Alejandro G. Villasante-Tezanos²

The University of Texas Medical Branch at Galveston

Research Article

Keywords: Artificial intelligence, Objective structured clinical examination, Standardized patients, Medical education assessment, Inter-rater reliability, Simulation-based education

Posted Date: September 28th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10767584/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Introduction

The Objective Structured Clinical Examination (OSCE) is widely used to assess clinical competence in medical education. Standardized Patients (SPs) often score learner performance, but this process is time-consuming, resource-intensive and susceptible to variability. As simulation platforms increasingly incorporate AI-assisted transcription and scoring features, medical education programs need evidence about what AI can and cannot reliably assess. This study evaluated AI scoring of objective medical interview checklist items by comparing AI-generated and SP scores using de-identified transcripts from 234 first-year medical students.

Methods

Recorded OSCE encounters were transcribed and scored using Insight7 with AI prompts developed from the 12 objective medical interview checklist items analyzed in this study. These items were selected because they could be evaluated using transcript-based evidence. Accuracy was assessed using descriptive statistics and Cohen's Kappa. Discrepancy analysis identified sources of scoring differences across the 12 analyzed items.

Results

AI demonstrated moderate inter-rater reliability with SP scoring for the 12 objective medical interview items analyzed (Kappa = 0.663). The average total score across these 12 items was 11.51 (SD = 0.78) for AI scoring, compared to 11.60 (SD = 0.71) for SP scoring. AI scoring required 2.75 hours, compared to 56.3 hours for SP scoring processes. Across 2,808 item-level scoring decisions (12 items × 234 encounters), 83 discrepancies (3%) were identified, primarily due to transcription limitations, varied student phrasing, nonverbal or implied responses, and differences in scoring judgment.

Conclusions

AI-assisted scoring showed moderate agreement with SP scoring for objective medical interview checklist items and substantially reduced scoring time. However, the findings also show that AI scoring remains dependent on transcript accuracy, prompt design, transparent scoring logic, and case-specific validation. These results support cautious use of AI as a scoring aid for structured checklist items, while reinforcing the need for human oversight and additional research before AI is used for interpersonal or relational skill assessment.

Introduction

Background/Rationale

Artificial intelligence (AI) is being explored to enhance medical education, particularly for its potential to support assessment, feedback, and documentation in clinical education.^{1–3} Evaluation of clinical competence during an Objective Structured Clinical Examination (OSCE) is commonly performed by faculty or standardized patients (SPs). This process is time-consuming, resource-intensive, and susceptible to variability.^{4,5} At the same time, commercial simulation platforms are beginning to incorporate AI-assisted evaluation features, including transcription, checklist scoring, SOAP note scoring, confidence indicators, and source-linked review tools.^{6,7} These developments increase the need for evidence about the conditions under which AI scoring is accurate, transparent, and appropriate for simulation-based assessment. This study examines AI scoring for objective medical interview checklist items and identifies practical limitations that remain important as AI-enabled scoring tools become more available.

The OSCE is widely regarded as the gold standard for evaluating clinical competence.^{8–10} First introduced by Harden et al. in 1975, it was designed to provide a structured, objective means of evaluating medical students across core competencies through SP encounters. In these encounters, students rotate through multiple stations, performing tasks such as interviewing a patient or performing physical exam maneuvers.¹¹ These evaluations are widely used to ensure that students demonstrate competency in real-world clinical scenarios. Despite its widespread adoption, challenges persist regarding consistency in scoring and feedback.^{12,13}

Initial research into AI's capacity to assist in scoring indicates that AI can identify structured, objective elements when guided by clear rubrics and carefully designed prompts. However, the human component remains critical for interpreting nuanced, context-dependent, or relational content. In areas such as post-encounter note evaluation, studies show promising results for AI-assisted tools when they are supported by human oversight and well-calibrated rubrics.³ These studies demonstrate AI's potential as a scalable tool in educational assessment while reinforcing the importance of human facilitation and interpretation. For that reason, the present study focuses on objective medical interview checklist items that could be evaluated from transcript content.

Standardized Patients in OSCE Scoring

SPs are trained individuals who consistently and reproducibly portray patients across a range of clinical conditions, including specific medical histories, symptoms, and emotional responses. Interactions with SPs allow students to practice and demonstrate clinical skills in a safe, realistic setting. The primary advantage of working with SPs is that they provide a consistent, realistic patient portrayal for teaching and evaluating clinical skills in medical education.^{5,14} SPs are also commonly involved in scoring observable learner behaviors during OSCE encounters, while faculty typically assess higher-order clinical competencies, such as clinical reasoning and decision-making.¹⁵

Despite their advantages, employing SPs to assess student performance also presents challenges. Training SPs to score student performance can be time-consuming and expensive. Maintaining timely and effective quality assurance necessitates regular training, monitoring, and evaluation of SPs.¹⁶ To achieve reliability, SPs must undergo extensive training, which includes reviewing the case and checklist, participating in mock encounters, and practicing checklist completion to achieve inter-rater consensus.¹⁷ Furthermore, despite extensive training, inconsistencies may arise in how SPs portray scenarios or assess students, potentially affecting the reliability of the scores. For example, some SPs are more lenient than others, and SP leniencies may vary depending on the clinical aspect being rated.¹⁶ Despite various grading strategies designed to reduce variability, human involvement will inevitably introduce some degree of variability and bias. The literature shows that the objectivity of the assessment can be compromised, as SPs tend to overscore medical students in the absence of a complete understanding of what is expected of them.¹⁸

Given these challenges, AI offers a possible way to improve efficiency and consistency for selected components of OSCE scoring, particularly checklist items that are objective and can be verified through speech captured in a transcript.^{19,20} Recent studies have explored the feasibility of using generative AI to automate OSCE scoring, with promising results when comparing AI-generated scores with those of human scorers.²¹ However, AI scoring should not be treated as inherently objective or interchangeable with human scoring. AI models may introduce different limitations, including dependence on transcript quality, prompt wording, model behavior, and whether the model was trained on data that reflects the learners, communication styles, and clinical simulation contexts being assessed. These limitations are especially important in healthcare education, where communication, empathy, professionalism, and patient-centered care remain central outcomes. AI may reduce some operational burdens, but its use in assessment must be transparent, validated, and limited to domains where the scoring evidence is appropriate.

AI's difficulty with the full nuance of human interaction raises important questions about its appropriateness for assessing relational competencies. Human scoring may be variable, but it can capture context, tone, nonverbal behavior, and interpersonal intent in ways that transcript-based AI cannot reliably reproduce.²² Therefore, this study does not position AI as a replacement for SP judgment. Instead, it evaluates whether AI can support scoring of objective medical interview checklist items while identifying limitations that must be addressed before AI scoring is expanded to broader or higher-stakes uses.

Objectives

This study evaluates AI-based scoring compared with SP scoring for objective medical interview checklist items in a first-year medical student OSCE. Specifically, the study has two objectives:

1. Investigate the accuracy of AI-assisted scoring compared with SP scoring using transcriptions of previously recorded medical interview simulations and SP-completed evaluation checklists.
2. Identify causes of discrepancies between AI and SP scoring by examining the influence of checklist adaptation, transcript accuracy, and scoring judgment on outcomes.

This study examined the level of agreement between AI-assisted and SP-driven scoring using a structured medical interview evaluation checklist. The scoring data included 234 first-year medical students participating in an OSCE. The OSCE format was chosen for its importance in medical education for evaluating students' skills and for addressing the inherent challenges of human grading variability. It was hypothesized that AI-assisted scoring would demonstrate moderate agreement with SP scoring when applied to objective OSCE checklist items.

Materials and Methods

Study Design

This study utilized a cross-sectional comparative design to evaluate the agreement between AI-generated and SP-assigned scores for a single OSCE station. Transcripts were generated retrospectively from recorded clinical encounters. SP-scoring was supported by established training and quality assurance procedures, including inter-rater reliability review. Additional QA review was prioritized for encounters near performance thresholds or otherwise requiring review, while selected encounters above the passing threshold were sampled to monitor scoring consistency.

Setting

The research took place at an academic health education center within its Standardized Patient Center, which regularly conducts OSCEs. The center employs trained SPs from the community who simulate various medical conditions, histories, and emotional responses to help students develop clinical and communication skills. SPs play a critical role in OSCEs by supporting consistent, reproducible learner encounters and completing structured evaluations through established training and quality assurance protocols.

Participants

Using previously recorded OSCE encounters conducted as part of the first-year medical curriculum in the school of medicine, the sample comprised 234 first-year medical students and six SPs who participated in this OSCE encounter from October 9 to October 20, 2023. These encounters were recorded as part of standard educational activities. The data relied on retrospective analysis of de-identified video recording transcripts; no new student recruitment was involved.

Ethical Approval

and Informed Consent

This study involved retrospective analysis of previously recorded educational data. The study protocol was reviewed by the authors' Institutional Review Board and was determined to be exempt from formal review as secondary analysis of de-identified educational data (IRB #24-0331). The requirement for informed consent was waived because the study involved secondary use of de-identified data collected as part of routine educational activities.

Procedures

The study was conducted between November 1, 2024, and February 15, 2025, and followed a two-phase approach. In the first phase, recorded OSCE videos were transcribed and anonymized into de-identified transcripts. Transcript quality was examined to support accurate AI scoring and to help explain discrepancies between AI and SP scores. In the second phase, Insight7 (insight7.io, Wilmington, DE), a transcription and qualitative data analysis software platform with AI-assisted features, was used to generate transcripts and AI-based scores.

The SP checklist items were initially entered as AI prompts with minimal modification. These prompts were then revised to improve scoring accuracy, as detailed later in the manuscript. Insight7 was used to develop AI prompts and execute the scoring process. Transcripts were created in Insight7 from recorded encounter videos, and the finalized prompts were applied to generate checklist scores. Because speaker labels (e.g., learner versus patient) were not consistently identified in the transcripts, the AI evaluated the entire conversation rather than isolating learner dialogue. As a result, the AI could assign credit based on either the student's question or the standardized patient's response. This differed from SP scoring, which generally prioritizes student questioning but may also consider the SP response when determining whether the learner elicited the required information.

The research team validated the AI scoring workflow by comparing AI-generated scores with SP scores. This included a review of discrepancies to assess adherence to rubric criteria and consistency across encounters. Inter-rater reliability was quantified using Cohen's Kappa, a statistical measure of agreement between AI and SP scoring. Validation was intended to examine scoring consistency and identify sources of disagreement. This study was not designed as an equivalence trial, and Cohen's Kappa was used to measure association rather than equivalence.

Discrepancy was defined as any difference between the AI-generated and SP-assigned scores for the same checklist item. Scoring error was defined as a deviation from rubric-based expectations on a checklist item. Scoring errors were identified using a manual review of transcript content and associated scores to confirm alignment between AI and SP scores and the rubric. The review process aimed to determine whether the score assigned by either AI or SP met the rubric criteria relative to the actual dialogue. This process did not serve as a formal adjudication to verify one individual's score versus the other, but rather as a rubric-based consistency check to understand scoring behavior and inform

discussion of limitations. Discrepant scores were reviewed to better understand and characterize scoring variances, consistent with best practices in iterative instrument improvement.

Transcription Accuracy Testing. To ensure transcription accuracy, transcription quality was evaluated using three transcription types: manual human-based transcription, Insight7, and Microsoft Stream. Three OSCE videos (Students 11, 49, and 50) representing high, medium, and low student performance were selected, manually transcribed by a researcher, and verified for accuracy. These transcriptions were compared to those generated by Insight7 and Microsoft Stream automated transcriptions. Accuracy was calculated by tallying word-level errors (e.g., missing or misspelled words). Insight7 achieved 99.4% accuracy, outperforming Microsoft Stream (94.7%). Based on this analysis, Insight7 was selected for its transcription fidelity, speed, and compatibility with AI scoring tools.

Estimated time differences between AI and SP scoring. The estimated time investment required for SP-based scoring versus AI-based scoring using the Insight7 platform highlights the resource-intensive nature of the traditional SP-scoring approach. The traditional approach included SP training (7 hours total) and encounter grading (80 minutes per SP per day × 10 days = 13.3 hours). In addition, simulation center staff conducted an inter-rater reliability review to support consistency and scoring accuracy. Although the time required for this process was not formally tracked and could not be reliably estimated, it constituted a necessary and ongoing layer of quality control in the SP-scoring workflow.

In contrast, the AI-based scoring process initially required approximately 2.75 hours for video uploading, automated transcription, and checklist scoring across all 234 student encounters. However, additional time was invested in preparing and validating the AI scoring workflow to ensure accuracy and alignment with expected performance standards.

For this study, analysis was limited to the 12 objective medical interview items because these items could be evaluated using transcript-based evidence. The 12 analyzed checklist items were refined through repeated testing and revision into rule-based prompts. Although the time spent on prompt development was not formally recorded, it is estimated to have required approximately 4 to 6 hours, based on the number of revisions and manual testing conducted during the calibration process.

When all components of the AI scoring workflow are considered, including video processing, prompt development, transcript validation, and manual benchmarking, the total time investment is estimated to be between 11 and 14.5 hours. Although this is greater than the 2.75 hours for AI to analyze the transcriptions and produce scores, it remains substantially lower than the estimated time required for traditional SP-based scoring processes. As such, the AI-based approach offers substantial long-term gains in efficiency and scalability for high-volume OSCE assessments.

Checklist Adaptation for AI Grading. For the present study, analysis focused on 12 objective medical interview items, including whether students asked about symptom onset, location, character, relief, radiation, severity, medical conditions, surgeries, medications, family risk factors, tobacco use, and diet. These checklist items were appropriate for transcript-based AI scoring because they were tied to

verbalized history-taking behaviors that could be identified in the encounter transcript. Each item was converted into an objective-based AI prompt aligned with common student phrasing observed in the transcripts.

This study, therefore, focused on the Medical Interview section (see Table 1 for the 12 analyzed checklist items) and converted each item into an objective-based prompt. These prompts were aligned with common student phrasing, enabling the AI to detect checklist items based solely on transcript content. This modification improved scoring reliability by aligning checklist design with natural-language processing capabilities.

Variables

This study compared AI and SP scoring of OSCE encounters by examining the accuracy of AI-generated scores and discrepancies between AI-generated and SP-assigned scores. Accuracy refers to the degree to which AI-generated scores aligned with SP scores that had undergone the usual quality assurance and inter-rater reliability processes. The analysis included 12 objective medical interview checklist items, item-level scores, total scores across the 12 analyzed items, and score differences between AI and SP scoring.

Statistical Analysis

Accuracy was examined using descriptive statistics (mean and standard deviation) for both AI and SP scorers, along with the average number of discrepancies. Although the data were not normally distributed, means and standard deviations were reported to facilitate interpretation. Given the limited variability, alternative measures such as interquartile ranges would not provide additional meaningful information. These descriptive statistics were then compared. In this study, accuracy refers to the degree to which AI-generated scores aligned with SP scores.²² To assess agreement between AI and SP scoring, inter-rater reliability was calculated using Cohen's Kappa for ordinal checklist ratings of 0, 0.5, or 1. A value less than 0.40 is considered slight agreement,²³ indicating a need to refine either the AI prompts or human-assessor training. Analysis was conducted in R 4.4.3. This study does not establish a threshold for acceptable agreement in summative evaluation. The reported Cohen's Kappa value of .663 reflects moderate agreement, consistent with published guidelines, and is presented descriptively rather than as a benchmark for sufficiency.²³

A qualitative review of 83 scoring discrepancies across 2,808 checklist items (12 items × 234 encounters) was conducted to identify and understand sources of disagreement. The first author coded the 83 discrepancies, providing a brief description of the reason for each, and the second and third authors reviewed the assigned codes. Any disagreements in the assigned code were resolved through discussion. From the codes, the first three authors identified broad categories that captured overarching patterns of discrepancy between AI and SP scoring.

Results

This investigation examined the 12 objective medical interview checklist items only. This narrower scope allowed the analysis to focus on whether AI could score checklist items that were clearly tied to verbalized history-taking behaviors.

Descriptive Statistics

Overall, after prompt refinement, AI demonstrated moderate agreement with SP scoring for the 12 objective medical interview checklist items. Of the 83 discrepancies identified out of 2,808 item-level scoring decisions, 59% were attributed to AI scoring errors, 37% to SP scoring errors, and 4% were ambiguous. Four key categories emerged in the analysis of these discrepancies: (1) transcription-related limitations, such as low audio volume, background interruptions, and overlapping speech, resulted in missing or misattributed content; (2) variation in how students phrased checklist-relevant content occasionally led the AI to miss correct responses when phrasing deviated from expected patterns; (3) nonverbal or implied communication, such as gestures or tone, was not captured in transcripts and therefore not recognized by the AI; and (4) differences in scoring judgment, where the AI applied strict text-matching logic, while SPs often credited implied intent or partial responses.

Accuracy

Across the 12 medical interview items analyzed, the average total score given by AI was 11.51 (SD = 0.78), and the average total score given by human scorers was 11.60 (SD = 0.71). Table 1 outlines means and standard deviations (SD) for AI and SP scoring across the 12 analyzed checklist items. AI scoring closely aligned with SP scores across most checklist items, particularly for highly structured questions such as pain onset and family history. Minor discrepancies in items like surgical history and tobacco use highlight areas where AI struggled to interpret varied phrasing or incomplete responses.

Table 1

Accuracy of scoring across the 12 analyzed objective medical interview checklist items by AI and SP

Checklist Item	SP Mean	SP SD	AI Mean	AI SD	Mean Difference	SD Difference
Onset	1	0	1	0	0	0
Location	0.99	0.09	0.95	0.21	0.04	-0.12
Character	0.99	0.11	0.97	0.16	0.01	-0.05
Relief	0.99	0.08	0.99	0.07	0	0.01
Radiation	0.94	0.23	0.94	0.24	0	-0.01
Severity	0.96	0.2	0.95	0.21	0.01	-0.02
Medical Conditions	0.97	0.16	0.99	0.11	-0.01	0.05
Surgeries	0.97	0.16	0.92	0.27	0.05	-0.11
Medications	0.97	0.15	0.95	0.17	0.01	-0.02
Family Risk Factors	1	0.03	1	0.03	0	0
Tobacco Use	0.87	0.25	0.89	0.24	-0.03	0.01
Diet	0.95	0.22	0.94	0.22	0.01	0

Inter-rater reliability was calculated from the individual item-level scores using Cohen's Kappa to assess agreement between AI and SP scoring. A Cohen's Kappa value of 0.663 was obtained, indicating moderate agreement with SP scoring, as defined by Landis and Koch.²³

Discrepancies

Discrepancies between AI and SP scores were primarily driven by transcription errors, ambiguous phrasing, and variation in how students verbally expressed key ideas, making it difficult for AI to interpret responses consistently. While SPs and AI used the same 12 objective medical interview checklist items, SPs occasionally inferred intent from vague or incomplete wording, awarding credit even when the response lacked precise language. In contrast, AI relied strictly on the explicit verbal content within the transcript. The AI was more likely to mark the item as incomplete or missing when a student did not clearly state or ask a required checklist item or used vague language. Scoring issues also occurred when dialogue overlapped, voice volume was low, or interruptions occurred, such as overhead announcements.

To further explore the cause of discrepancies, qualitative review revealed four key categories: (1) transcription-related limitations (e.g., low volume, background interruptions, and overlapping speech); (2) variation in how students phrased checklist-relevant content, which affected AI detection; (3) nonverbal or implied responses not captured in text; and (4) scoring judgment differences, with SPs sometimes awarding credit based on inferred intent while AI required explicit statements. The four categories are

summarized in Table 2, which presents representative examples illustrating how each discrepancy type affected scoring accuracy.

Table 2
Themes related to discrepancies between AI and SP scoring

Theme	Definition	Representative Example
Transcription-Related Limitations	Technical issues in the transcripts, such as low voice volume, overlapping speech, background interruptions (e.g., time-warning announcements), and misheard or missing words, prevented AI from accurately identifying checklist items.	Misheard/missing words: The word "resonate" was dictated in the transcript instead of "radiate." AI cannot interpret incorrect word usage. Student: "Does the pain resonate anywhere?" Patient: "No, it stays right here."
Variation in Student Language and Phrasing	Students used diverse language, multipart questions, and indirect phrasing to address checklist items. SPs may give credit based on intent or inferred context, while AI requires explicit, transcript-based evidence.	Diverse language and inferred meaning: AI may fail to award credit when valid student phrasing does not match the wording anticipated by the prompt. Student: "Do you feel the pain anywhere else on your body?" Patient: "No, it stays right here."
Nonverbal and Implied Responses	SPs occasionally award credit based on their own nonverbal responses (e.g., nods, tone, facial expressions) or implied meaning. These cues are invisible to the AI, which relies solely on the transcript.	Nonverbal SP responses: When SPs respond nonverbally, such as nodding or pointing, AI cannot detect these actions in the transcript, leading to scoring errors. Student: "Can you point to where the pain is located?" Patient: "It's right here (points to pain location)."
Scoring Judgment Differences	Scoring discrepancies may arise when SPs incorrectly award or deny credit even though the student asked the appropriate question and received the correct response. Such inconsistencies reflect variability in SP scoring accuracy. Alternatively, the AI may fail to award credit due to overly rigid prompt parameters, limiting its ability to recognize paraphrased questions or answers.	AI Prompt Too Rigid or Narrow: In some cases, AI scoring may fail to award credit when a student uses synonymous or related terms instead of the exact phrasing anticipated by the AI prompts. This limitation likely arises from the AI's reliance on narrow or rigid prompt parameters, which reduces its ability to recognize valid variations in student responses. The example below illustrates how the AI expected specific terminology not reflected in the student's question or the standardized patient's reply. Student: "You mentioned a prescription medication earlier. Are you taking other medications or supplements, like over-the-counter products or herbal remedies?" Patient: "Yes, I do take a daily supplement."

Additionally, early versions of the AI prompts lacked the clarity and flexibility needed to align with the varied ways students phrased clinical questions. Some prompts were too narrowly worded and failed to recognize acceptable responses unless students used precise phrasing. Others were overly broad, introducing ambiguity that could lead to credit for responses that were only partially relevant or off topic.

The 12 objective items from the Medical Interview section were converted into objective-based AI prompts aligned with common student language patterns observed in the transcripts. This restructuring, conducted within the Insight7 platform, improved AI scoring consistency by enhancing the prompts' ability to detect checklist elements in natural language while preserving alignment with SP criteria for the analyzed medical interview items.

Discussion

This study found that AI-assisted scoring showed moderate agreement with SP scoring for the 12 analyzed objective medical interview checklist items in a first-year medical student OSCE. Across those 12 items, the average total score given by AI was 11.51 (SD = 0.78), compared with 11.60 (SD = 0.71) for SP scorers. AI completed the scoring process in substantially less time than the SP-based scoring workflow, highlighting its potential value for high-volume, structured assessments. However, the findings also show that AI scoring is not simply an automated replacement for human evaluation. Discrepancies were primarily related to transcription limitations, variation in student phrasing, nonverbal or implied responses, and differences in scoring judgment. These issues remained important even within the narrow scope of objective medical interview items. Therefore, the value of AI scoring is best understood as a potential scoring aid for carefully selected checklist items, not as a broad substitute for SP or faculty judgment.

These findings are especially relevant now because AI-assisted evaluation is moving from experimental use into commercial simulation platforms. Publicly available product information for LearningSpace describes AI-assisted transcription, checklist scoring, SOAP evaluation support, confidence indicators, and source-linked review features, while also emphasizing that faculty remain in control of final evaluation decisions.^{6,7} This direction supports the need for studies like the present one. Even when software can recommend scores, unresolved questions remain about prompt design, scoring transparency, case-specific validation, transcript quality, and whether AI-generated scores should be trusted without human review. The present study contributes practical evidence to that conversation by showing both the promise and the limitations of AI scoring under controlled, retrospective conditions.

Several limitations should be considered. This study examined one OSCE station at a single institution, which limits generalizability to other cases, disciplines, learner levels, and assessment formats. AI scoring was also dependent on the quality of the transcript. Poor audio quality, overlapping speech, interruptions, and missing or misattributed words can affect whether AI recognizes a checklist item. In addition, the prompts were developed and refined using retrospective transcripts and known SP scores. This means the scoring workflow benefited from prior data that would not be available when launching a

new case for the first time. Future implementations should begin with human scoring and retrospective validation before AI scoring is used operationally or in any higher-stakes setting.

Furthermore, the Insight7 software used in this study employs large language models to analyze transcript data and produce scoring outputs in response to predefined prompts. Because the specific architecture and training dataset for Insight7 are proprietary and not publicly disclosed in full, complete transparency into the model's training data is not available. As with any AI system, this limits the ability to fully assess model bias, including potential limitations related to culturally specific language, underrepresented communication styles, or variations in how learners phrase clinical questions. Future studies should consider working with platforms that enable greater transparency or independently validating model outputs across diverse learner populations.

Despite AI scoring limitations, AI performed well when using structured transcripts and rule-based prompts. This aligns with prior research on AI grading of post-encounter notes, which similarly emphasized the need for human judgment to guide interpretation and ensure scoring validity.³ It also aligns with previous research indicating that multiple large language models (LLMs) can accurately score OSCE encounters.^{21,24} This study builds on this work by extending AI evaluation to OSCE checklist scoring, broadening its application in simulation-based medical education, and demonstrating its feasibility for structured clinical performance assessments.

The discrepancies outlined in Table 2 reveal distinct patterns in how AI and SP scoring diverge. These categories offered insight into the operational differences between AI scoring logic and SP judgment, underscoring opportunities for prompt refinement and SP training. Most AI scoring errors stemmed from rigid interpretation of checklist prompts, with students using paraphrased or indirect phrasing that the AI failed to recognize as correct. In contrast, SPs sometimes over- or under-credited behaviors by inferring intent from context or implied meaning rather than relying solely on the checklist language. A smaller number of discrepancies involved ambiguous situations or transcription limitations, such as missing audio or nonverbal responses. These findings reinforce the importance of refining AI prompt logic to handle natural language variation while ensuring consistent SP training to reduce subjective interpretation. Including transcript excerpts alongside each discrepancy enhanced transparency by allowing researchers to assess how closely AI and SP interpretations aligned with the actual dialogue from the encounter. Together, these findings underscore the complementary limitations of each method and the opportunity for targeted improvements in both human and AI-supported assessment.

AI deployment should include precise prompts, consistent transcription, ongoing error analysis, and a clear process for human review. Institutions seeking to adopt AI-based OSCE scoring should begin with retrospective data to calibrate prompts and validate scoring logic before implementation. They should also require transparency about how scores are generated, what evidence is used, how confidence is reported, and how faculty or SPs can review the source material. Discrepancy analyses like those in Table 2 can serve as a useful quality assurance process because they show where AI scoring fails,

where SP scoring varies, and where the checklist itself may need clearer scoring criteria. These steps are necessary if AI scoring is to be used safely and responsibly in clinical skills assessment.

While this study demonstrates that AI scoring can be accurate when applied to retrospectively calibrated data, it also highlights a key challenge in applying AI to new, untested simulations. Each AI scoring prompt in this study was iteratively developed using previously recorded transcripts and known SP scores. This process allowed the research team to calibrate prompts, resolve ambiguity, and ensure alignment with common student phrasing. However, such data would not yet exist in a new case scenario. These findings align with recent work on prompt engineering in medical education, emphasizing the need for careful prompt construction to ensure reliable AI output.²⁵ Without baseline human scores and accurate transcriptions, developers cannot fully anticipate how students will phrase questions or which checklist items will trigger errors. Without case-specific calibration, the use of AI scoring carries significant risk. Consequently, human scoring is essential during the initial testing of new simulation cases. Only after iterative validation can AI be reliably applied in those contexts.

To support widespread adoption, future studies should replicate these findings across multiple institutions and simulation formats, using varied learner populations. Additionally, future research could build on these findings by evaluating the actual labor costs and financial implications of AI versus SP scoring beyond time-based estimates.

Future research should also examine how AI-enabled simulation platforms perform when scoring new cases prospectively rather than retrospectively. Important questions remain about whether commercial tools use institution-specific prompts, embedded scoring criteria, automated confidence ratings, or human-in-the-loop review processes, and whether these approaches resolve the limitations identified in this study. Additional research is also needed to determine whether AI can validly support assessment of interpersonal and relational skills using appropriate evidence sources, multimodal data, and human oversight. Until those questions are answered, AI scoring should be limited to clearly defined, objective checklist items and used as part of a supervised assessment workflow.

Conclusions

AI-assisted scoring demonstrated moderate agreement with SP scoring for 12 objective medical interview checklist items and substantially reduced scoring time in this OSCE setting. However, the findings also show that AI scoring depends on accurate transcription, clear prompt design, case-specific validation, and human oversight. AI-based scoring may be useful as a supervised support tool for structured checklist items, but it should not be implemented for higher-stakes assessment or broader competency domains without careful validation.

Declarations

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Ethics Approval

This study involved retrospective analysis of previously recorded educational data. The study protocol was reviewed by the University of Texas Medical Branch Institutional Review Board and was determined to be exempt from formal review as secondary analysis of de-identified educational data (Institutional Review Board [IRB #24-0331]).

Consent to Participate

Informed consent was waived due to the retrospective use of de-identified data collected as part of routine educational activities.

Consent for Publication

Not applicable.

AI Usage Statement

Artificial intelligence was used as part of the study methodology to analyze transcribed OSCE encounters and generate scoring outputs using the Insight7 platform. AI was also used to assist with language refinement during manuscript preparation. All AI-assisted outputs were reviewed and verified by the authors.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contribution

Richard L. Briley contributed to study design, data analysis, AI prompt development, discrepancy coding, and manuscript drafting. Kelly R. Grace contributed to assessment design, discrepancy review, and manuscript editing. Bruce W. Adcock contributed to discrepancy review, methodological refinement, and manuscript editing. Erin C. King, Amanda D. Carey, Erin K. McGoff, and Patricia D. Self contributed to simulation implementation, data collection oversight, and manuscript review. Alejandro G. Villasante-Tezanos contributed to statistical analysis and manuscript review. All authors reviewed and approved the final manuscript.

Data Availability

The datasets generated and analyzed during the current study are not publicly available due to institutional and educational privacy restrictions but may be available from the corresponding author on reasonable request and with institutional approval.

References

1. Misra SM, Suresh S. Artificial Intelligence and Objective Structured Clinical Examinations: Using ChatGPT to Revolutionize Clinical Skills Assessment in Medical Education. *J Med Educ Curric Dev*. 2024;11:1–6. 10.1177/23821205241263475.
2. Liu A, Martin S, Belousov M, Fung CC, Ben-Ari R, May W. Impact of AI Virtual Standardized Patients (AI-VSP) on performance in a Clinical Skills Examination. *Research Square*. Preprint posted online November. 2023;1. 10.21203/rs.3.rs-3516363/v1.
3. Jamieson AR, Holcomb MJ, Dalton TO, et al. Rubrics to Prompts: Assessing Medical Student Post-Encounter Notes with AI. *NEJM AI*. 2024;1(12):Alcs2400631. 10.1056/Alcs2400631.
4. Glassman PA, Luck J, O’Gara EM, Peabody JW. Using Standardized Patients to Measure Quality: Evidence from the Literature and a Prospective Study. *Jt Comm J Qual Improv*. 2000;26(11):644–53. 10.1016/S1070-3241(00)26055-0.
5. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *AAMC. Acad Med J Assoc Am Med Coll*. 1993;68(6):443–451; discussion 451–453. 10.1097/00001888-199306000-00002
6. Elevate Healthcare. LearningSpace: Simulation Management & Debriefing Platform. Accessed August 20. 2026. <https://elevatehealth.net/product/learningspace/>
7. Elevate Healthcare. LearningSpace AI – Ehli. Accessed August 20. 2026. <https://elevatehealth.net/learningspace-ai-ehli/>
8. 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*. Published online May. 2023;4. 10.7759/cureus.38519.
9. Norman G. Research in medical education: three decades of progress. *BMJ*. 2002;324(7353):1560–2. 10.1136/bmj.324.7353.1560.
10. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. *Ann Surg*. 1995;222(6):735–42. 10.1097/00000658-199512000-00007.
11. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *BMJ*. 1975;1(5955):447–51. 10.1136/bmj.1.5955.447.
12. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores: Reliability of objective structured clinical examination

- scores. *Med Educ.* 2011;45(12):1181–9. 10.1111/j.1365-2923.2011.04075.x.
13. Wong WYA, Roberts C, Thistlethwaite J. Impact of Structured Feedback on Examiner Judgements in Objective Structured Clinical Examinations (OSCEs) Using Generalisability Theory. *Health Prof Educ.* 2020;6(2):271–81. 10.1016/j.hpe.2020.02.005.
 14. May W, Park JH, Lee JP. A ten-year review of the literature on the use of standardized patients in teaching and learning: 1996–2005. *Med Teach.* 2009;31(6):487–92. 10.1080/01421590802530898.
 15. Flanagan OL, Cummings KM. Standardized Patients in Medical Education: A Review of the Literature. *Cureus* Published online July. 2023;17. 10.7759/cureus.42027.
 16. Brown CB, Kahraman N. Exploring Psychometric Models to Enhance Standardized Patient Quality Assurance: Evaluating Standardized Patient Performance Over Time. *Acad Med.* 2013;88(6):866. 10.1097/ACM.0b013e3182901647.
 17. Erby LAH, Roter DL, Biesecker BB. Examination of standardized patient performance: Accuracy and consistency of six standardized patients over time. *Patient Educ Couns.* 2011;85(2):194–200. 10.1016/j.pec.2010.10.005.
 18. Wilkinson TJ, Fontaine S. Patients' global ratings of student competence. Unreliable contamination or gold standard? Patients' ratings of student competence. *Med Educ.* 2002;36(12):1117–21. 10.1046/j.1365-2923.2002.01379.x.
 19. Smith A, Pietersen C, Van Wyk L, Van Wyk BJ. Grading variability in software design assessments: Human versus artificial intelligence. In: Collier-Reed B, ed. *Proceedings of the Sixth Biennial Conference of the South African Society for Engineering Education.* 2023:231–243.
 20. Campbell KK, Holcomb MJ, Vedovato S, et al. Applying State-of-the-Art Artificial Intelligence to Simulation-based Education: A Study of Comparative Educational and Operational Benefits. Preprint posted online July. 2024;18. 10.21203/rs.3.rs-4572735/v1.
 21. Geathers J, Hicke Y, Chan C et al. Benchmarking Generative AI for Scoring Medical Student Interviews in Objective Structured Clinical Examinations (OSCEs). *arXiv.* Preprint posted online 2025. 10.48550/ARXIV.2501.13957
 22. East L, Stevens A, Hageman H, et al. Validity evidence for a feedback tool for an interprofessional standardized patient experience: A qualitative pilot study with evaluation of content, response processes, consequences of testing, and feedback quality. *J Interprofessional Educ Pract.* 2025;40:100749. 10.1016/j.xjep.2025.100749.
 23. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics.* 1977;33(1):159. 10.2307/2529310.
 24. Shakur AH, Holcomb MJ, Hein D et al. Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis. *arXiv.* Preprint posted online 2024. 10.48550/ARXIV.2410.12858
 25. Heston T, Khun C. Prompt Engineering in Medical Education. *Int Med Educ.* 2023;2(3):198–205. 10.3390/ime2030019.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AIScoringTable1.docx](#)
- [AIScoringTable2.docx](#)