BMC Medical Education

RESEARCH

Open Access

Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination

Murat Tekin[1], Mustafa Onur Yurdal[1], Çetin Toraman[1], Güneş Korkmaz[2,4*] and İbrahim Uysal[3]

Abstract

Background Objective Structured Clinical Examinations (OSCEs) are widely used in medical education to assess students' clinical and professional skills. Recent advancements in artificial intelligence (AI) offer opportunities to complement human evaluations. This study aims to explore the consistency between human and AI evaluators in assessing medical students' clinical skills during OSCE.

Methods This cross-sectional study was conducted at a state university in Turkey, focusing on pre-clinical medical students (Years 1, 2, and 3). Four clinical skills—intramuscular injection, square knot tying, basic life support, and urinary catheterization—were evaluated during OSCE at the end of the 2023–2024 academic year. Video recordings of the students' performances were assessed by five evaluators: a real-time human assessor, two video-based expert human assessors, and two AI-based systems (ChatGPT-4o and Gemini Flash 1.5). The evaluations were based on standardized checklists validated by the university. Data were collected from 196 students, with sample sizes ranging from 43 to 58 for each skill. Consistency among evaluators was analyzed using statistical methods.

Results AI models consistently assigned higher scores than human evaluators across all skills. For intramuscular injection, the mean total score given by AI was 28.23, while human evaluators averaged 25.25. For knot tying, AI scores averaged 16.07 versus 10.44 for humans. In basic life support, AI scores were 17.05 versus 16.48 for humans. For urinary catheterization, mean scores were similar (AI: 26.68; humans: 27.02), but showed considerable variance in individual criteria. Inter-rater consistency was higher for visually observable steps, while auditory tasks led to greater discrepancies between AI and human evaluators.

Conclusions AI shows promise as a supplemental tool for OSCE evaluation, especially for visually based clinical skills. However, its reliability varies depending on the perceptual demands of the skill being assessed. The higher and more uniform scores given by AI suggest potential for standardization, yet refinement is needed for accurate assessment of skills requiring verbal communication or auditory cues.

Keywords OSCE, Clinical skills assessment, Artificial intelligence, Medical education, Evaluator consistency, Interrater reliability

*Correspondence:
Güneş Korkmaz
gunes.korkmaz@medeniyet.edu.tr
[1]Medical Education, Çanakkale Onsekiz Mart University, Çanakkale, Turkey
[2]Medical Education, İstanbul Medeniyet University, İstanbul, Turkey
[3]Vocational School of Health Services, Çanakkale Onsekiz Mart University, Çanakkale, Turkey
[4]Faculty of Medicine, Department of Medical Education, İstanbul Medeniyet University, İstanbul, Turkey

Background

The rapid development of artificial intelligence (AI) has led to a growing presence of AI tools in educational settings, including medical education. A majority of Generation Z students appear to be more familiar and comfortable with integrating AI into their learning processes, often leveraging its benefits for productivity, personalization, and efficiency. However, learners and educators across all generations increasingly recognize both the potential and the limitations of AI use in education. While many value its usefulness in enhancing learning and reducing workload, concerns persist about over-reliance, ethical boundaries, and pedagogical appropriateness. As AI becomes more embedded in academic and professional environments, including clinical education, understanding its role in assessment remains a critical area of exploration [1–3].

Objective Structured Clinical Examination (OSCE), used to assess medical students' clinical skills, is an important assessment tool that allows educators to reliably evaluate students' practical and procedural skills, such as examining, diagnosing, and explaining a management plan for each case within a set time frame. These exams typically take place at different stations, with each station testing a different clinical skill. Despite challenges such as cost and sustainability, OSCEs have become widely used since their introduction in 1975 [4, 5]. In faculties with limited resources, particularly when there is a shortage of educators, students may miss out on this critical assessment tool, leading to significant differences in competition levels and student quality [6–9]. Cusimano et al. [10] suggested that decreasing budgets and the competitive nature of existing budgets have been two key factors forcing medical educators to rethink the application of OSCEs. In Turkey, another factor that medical schools hesitate to use OSCEs as a part of measurement and assessment procedure, is the issue of time costs, as faculty members simultaneously provide healthcare service to the society while teaching students and conduct scientific research. Given the current global economic crisis, the changing student profile, and the advancements in AI technologies, using current AI models as OSCE evaluators could reduce costs in medical schools, as well as accelerate student quality and technological transformation in education. Therefore, we believe it is necessary to test what contributions AI tools could offer in addressing a range of issues, particularly in medical schools where time costs are significant.

Although OSCEs have some limitations, they can be further enhanced with new technologies like AI to better evaluate and prepare medical students for their future careers [11]. With technological advancements, and considering the ever-changing demands of clinicians, AI could be integrated with complementary skills rather than resisting to change, capitalizing on both its strengths and weaknesses.

Advances in AI systems, particularly with Large Language Models (LLMs), have achieved superior performance in enhancing text-based AI tools, granting them human-like decision-making and reasoning capabilities. Simultaneously, there is a growing research trend focusing on extending these LLM-powered AI tools into the multimodal realm [12]. Multimodal Large Language Models (M-LLMs), which are AI systems trained on multiple data modalities, such as image, text, and audio, can process both textual and visual data together, allowing them to perform more complex and diverse tasks compared to text-only models. M-LLMs have made exciting progress as they enable AI agents to interpret and respond to various multimodal user queries, allowing them to perform more complex and nuanced tasks [13]. For example, models like GPT-4 V can write stories from images and perform mathematical calculations without requiring optical character recognition (OCR) [14].

In this context, M-LLMs can be used for personalizing educational content, assessing student performance, and creating interactive learning environments for learners. For example, through M-LLMs, intelligent teaching assistants that provide personalized feedback by analyzing student interactions or course content enriched with images could be developed [14]. In healthcare, M-LLMs can be used to analyze medical images, make diagnoses, and suggest treatment plans. Furthermore, these models can offer more accurate and comprehensive healthcare services by integrating medical text and images. For instance, models like LLaVA-Med can make clinical decisions supported by medical images and analyze patient reports more effectively than humans [13–15]. Therefore, M-LLMs should be considered as powerful AI tools offering significant opportunities in fields like education and healthcare. Thanks to their ability to process various types of data, they can leverage richer and more complex information sources to offer more effective and innovative solutions.

Traditionally, OSCE evaluations rely on human assessors, who must apply a set of predetermined criteria to each student's performance. While effective, human evaluations can be subject to inconsistencies, biases, and limited capacity for immediate feedback. AI, on the other hand, offers the promise of objective, real-time analysis with the ability to consistently apply evaluation criteria across all students. In our study, each student's OSCE performance will be evaluated by three expert human evaluators. One human evaluator assessed the student's live performance during the OSCE, while the other two evaluated the student's recorded performance. Additionally, AI-based multimodal language models, such as ChatGPT-4o and Gemini Flash 1.5, evaluated the same

procedural skills, and the performance of human evaluators was compared with AI-based assessments. The results of this study are expected to help us better understand the potential use of AI in medical education and contribute to improving future evaluation processes. The research questions formulated for this study are as follows:

1. Is there a significant difference in the consistency between AI-based multimodal language models (ChatGPT-4o and Gemini Flash 1.5) and human evaluators' assessments?
2. How do perception types (visual, auditory, and visual + auditory) influence the consistency between AI (ChatGPT-4o and Gemini Flash 1.5) and human evaluations of procedural clinical skills?

Methods

The study was conducted at a state university in Turkey to assess the clinical/practical skills acquired early in medical school (Years 1, 2, and 3) during an Objective Structured Clinical Examination (OSCE) for four skills: intramuscular (IM) injection, square knot tying, basic life support, and urinary catheterization. The research is a cross-sectional study examining the consistency between evaluations made by human evaluators during the OSCE, two expert reviewers after the exam, and two artificial intelligence (AI) evaluators after the exam. In line with the second research question of our study, the Bland-Altman analysis were conducted. To do this, each checklist criterion for the four procedural skills was categorized by perception type to better understand how sensory modality influenced the agreement between human and AI evaluations. Based on the nature of the skill components, criteria were classified into three categories:

- Visual (V): Steps that could be assessed based solely on visual observation, such as IM injection, tying a knot, or positioning equipment.
- Auditory (A): Steps that require verbal communication from the student, such as explaining the procedure to the patient or verbally confirming consent.
- Visual + Auditory (V + A): Steps that involve both visual action and verbal expression, such as introducing oneself while making eye contact or simultaneously performing and explaining a task.

Participants

The research data were obtained from 196 students (First, second, and third-year medical students from a state university in Turkey) who voluntarily agreed to participate by signing the informed consent forms allowing their performance to be video-recorded during the application of four specific skills: IM injection, tying a square knot, basic life support, and urinary catheterization in an Objective Structured Clinical Examination (OSCE) at the end of the 2023–2024 Academic Year (June 2024). This exam is a part of successfully completing their respective academic year and involves assessing the procedural skills they acquired throughout the year. The video recordings captured the performances of 43 students for the IM injection skill, 58 students for the square knot-tying skill, 47 students for basic life support, and 48 students for urinary catheterization.

Data collection tool

The data collection tool in this study consists of video recordings of medical students performing the skills in OSCE. For scoring, separate checklists were used for each skill. These checklists, prepared by the university to evaluate professional skills, are the ones which were published on the university's official website since 2018 and are used annually to assess students. When first developed, the checklists were created by a committee of expert physicians specializing in the relevant fields. They were then reviewed by another group of specialists and finalized under the supervision of a faculty member who is an expert in medical education with expertise in measurement and evaluation. Following this process, the checklists were officially published on the faculty's website, ensuring that students are informed in advance about the criteria by which their performance will be evaluated during practical exams. Since 2018, these checklists have been continuously updated based on feedback from students and evaluators during professional skill practices and OSCE sessions, thereby enhancing their validity and reliability.

Evaluators

Five different evaluators were involved in this study. The first evaluator assessed the students in real time during the OSCE and incorporated the scores into the year-end grades used for passing the class. The second evaluator, a specialist physician who participated in students' professional skills training during the academic year, evaluated the students based on video recordings. The third evaluator, another specialist physician who is not involved in the students' professional skills training during the academic year, evaluated the skills in the video recordings. The fourth evaluator was ChatGPT-4o, which evaluated the students based on video recordings. The fifth evaluator was Gemini Flash 1.5, which also evaluated the students from the video recordings.

Procedures

In this study, four clinical skills—intramuscular (IM) injection, square knot tying, basic life support (BLS), and

urinary catheterization—were selected for assessment. These skills were chosen based on their alignment with the university's professional skills curriculum for Years 1 to 3 of medical school, their fundamental role in early clinical training, and the availability of validated and widely used checklist-based evaluation tools. The selection was made by a committee of medical educators and clinical instructors who reviewed the curricular content and considered the feasibility of video-based and AI-based evaluation for each skill.

Prior to the OSCE, students were informed about the study and invited to participate voluntarily. Informed consent was obtained from those who agreed to be video recorded during the performance of the specified clinical skills. The OSCE was conducted as part of the regular academic year-end evaluation. For each consenting student, the performance at the corresponding OSCE station was recorded using a fixed camera setup that mirrored the perspective of the real-time evaluator.

Student performances were evaluated in five ways: by a human evaluator during the live OSCE, by two additional human experts using the video recordings, and by two AI systems (ChatGPT-4o and Gemini Flash 1.5) that assessed the same recordings based on standardized checklists. The checklists used for evaluation were developed by a panel of subject matter experts, reviewed for content validity. Each video was independently evaluated by the AI models without any prior training or feedback, using identical Turkish-language prompts and scoring rubrics.

Steps and Criteria for AI Evaluation of Video Recordings

1. *Model Selection*: During the research period, LLM-based models capable of handling video file sharing, specifically ChatGPT-4o and Gemini Flash 1.5, were selected for evaluation.
2. *Video Recording Specifications*: Videos were recorded from the same perspective and distance as the real-time human evaluator, ensuring a consistent viewpoint. Recordings were captured in 1920 × 1080 resolution at 30 FPS.
3. *Model Configuration*: Untrained models were used without fine-tuning. No feedback was provided to the AI models regarding their evaluation results.
4. Instructions to AI Models: Both AI models were provided with the evaluation form and scoring system only. They were instructed to assess the video recordings based on these forms.
5. *Prompt Details*: Identical prompts were input into both AI models in Turkish:

Prompt:*"Hello, we are conducting an OSCE exam in the Faculty of Medicine where students perform the "Urinary Catheterization Skill" on a mannequin. Evaluation criteria and scoring were prepared by expert physicians. We recorded videos of students performing the "Urinary Catheterization Skill" on mannequins. I will send you these videos. Could you evaluate the student's performance in the video, considering the verbal cues in the video and following the 15-step evaluation criteria provided? First, let me share the scoring system with you: If the student completes the step on time and fully, award 2 points. If the student performs the step hesitantly or incompletely, award 1 point. If the student does not perform the step at all, award 0 points."*

6. *Subsequent Evaluations*: For subsequent students, only a new prompt indicating the upload of a different video was entered.

Prompt: *"The performance video of the first student, Ha** Se*** AL***, will now be evaluated."*

7. *Evaluation Records*: Screenshots of the evaluations conducted by ChatGPT-4o and Gemini Flash 1.5 can be accessed **here.**

Data analysis

In this study, inter-rater reliability was examined using Krippendorff's Alpha and Fleiss Kappa inter-rater reliability analyses. Cohen's kappa coefficient (κ) is a statistical measure used to evaluate the consistency of ratings between two raters [16]. The κ statistic, initially introduced by Cohen as an indicator of agreement between two raters, was later adapted by Fleiss [17] to measure agreement among more than two raters [17]. According to the Fleiss Kappa coefficient, agreement levels are categorized as follows: 0.01–0.20: Negligible, 0.21–0.40: Poor, 0.41–0.60: Moderate, 0.61–0.80: Good, 0.81–1.00: Very good agreement [18].

Krippendorff's alpha is a reliability coefficient designed to assess the degree of agreement among observers, coders, judges, raters, or measurement tools when distinguishing between typically unstructured phenomena or assigning quantifiable values to them. Originally developed for content analysis, it is now broadly applicable in situations where multiple data generation methods are used on the same set of objects, units of analysis, or items. The focus of this method is on determining how reliable the resulting data is in accurately representing something real [19, 20].

Additionally, the evaluations made by both human and AI raters, considering all criteria for the relevant skill, were converted into a total score for each student. Three different comparisons were made using the obtained total scores. First, the evaluations of each rater regarding the students' skills were compared using a comparison

analysis (One Way ANOVA). Second, the evaluations of two human raters, who assessed the students via video recordings, were compared with those of a human rater who conducted the evaluation during the exam and the AI raters (One Way ANOVA). Third, human raters and AI raters were grouped separately, and a comparison (Independent Sample T-Test) was made between these two groups.

In addition to Fleiss Kappa and Cohen's Kappa analyses, Bland-Altman analysis was performed, and plots were generated in this study. Fleiss Kappa and Cohen's Kappa analyses are statistical methods used to measure the consistency of categorical data among multiple raters [16, 17]. Therefore, in this study, the evaluations made by individual human experts and AI models were treated as independent raters. To accurately assess the consistency between the evaluations, the individual scores of each rater were used instead of the average scores of the experts. This approach provides an objective and detailed representation of the level of agreement among the raters.

The rationale for applying the Bland-Altman analysis in this study lies in evaluating the consistency of two different measurement methods [21, 22]. Bland-Altman analysis is a reliable method developed to determine the systematic bias and the limits of agreement (LOA) between two measurement methods for continuous data. In our study, the average of the evaluations made by the three human raters for the four different clinical skills being assessed was considered a "gold standard" representing the general trend of human expertise, minimizing individual expert differences [23]. This average value was taken as the first measurement method, while the evaluations made by the AI models were considered the second measurement method. Using the average of human evaluations helps balance individual variations, providing a more reliable comparison. Furthermore, since both evaluation methods generate continuous data, Bland-Altman analysis was considered one of the most scientifically appropriate approaches to determine the consistency and possible systematic differences between these two methods. The results obtained from this method provide critical insights into comparing the performance of AI models with human expert evaluations and identifying any inconsistencies.

Ethics committee approval

This study was conducted with the approval of the Çanakkale Onsekiz Mart University Non-Interventional Clinical Research Ethics Committee (Date of Approval: 03.06.2024/No:2024/06–08).

Results

Intramuscular injection skill assessment

The assessment results of 43 students who participated in the intramuscular injection skill exam were analyzed using Krippendorff's Alpha and Fleiss' Kappa coefficients to assess the consistency between the in-exam evaluator, two human evaluators reviewing video recordings, and the evaluations conducted by ChatGPT-4o and Gemini Flash 1.5. Additionally, the arithmetic mean and standard deviation of the assessments conducted by the evaluators for the 15 criteria related to the intramuscular injection skill across the 43 students were also analyzed. The results are given in Table 1.

When examining the Krippendorff's Alpha values presented in Table 1, inter-rater consistency among the three human evaluators and two AI evaluators was not achieved across all 15 criteria for the intramuscular injection skill. The highest consistency was observed for the criterion "ensured privacy (verbal confirmation is sufficient)", while the lowest consistency was found for "inserted the needle perpendicular to the skin." Similarly, an analysis of Fleiss' Kappa values, calculated as an indicator of inter-rater agreement, showed the highest consistency for the criterion "informed the patient that the procedure was complete and/or said, 'Get well soon.'" The lowest consistency was recorded for "cleaned the injection area with an antiseptic cotton swab, wiping outward in a circular motion from the center, covering a radius of approximately 5 cm."

In nearly all criteria, human evaluators assigned lower scores compared to AI evaluators. Additionally, the standard deviation of scores provided by human evaluators was higher than that of AI evaluators, suggesting greater variability in the scores given by human evaluators. The total scores for the 15 criteria of the IM injection skill were calculated for each student, based on assessments by both human and AI evaluators. Subsequently, three types of comparisons were made: (1) A comparative analysis was conducted for the evaluations of each evaluator regarding the students' skills, (2) The assessments by the two human evaluators who reviewed the video recordings were compared with those of the human evaluator who conducted the exam and the AI evaluators, (3) The human evaluators were grouped as one category, and the AI evaluators as another, and a comparison was made between the two groups. The results of these comparisons are presented in Table 2.

Table 2 reveals that human evaluators provided significantly lower scores compared to AI evaluators ($p < .05$).

Square knot tying skill assessment

The assessment results for 58 students, evaluated by an examiner during the exam for the square knot skill, were analyzed alongside the results of two human evaluators

Table 1 Mean, standard deviation values, and Inter-Rater reliability level of evaluators

IM Injection Criteria	Inter-rater Reliability		Evaluator				
			H1	H2	RT	AI1	AI2
	α	κ	M(Sd)	M(Sd)	M(Sd)	M(Sd)	M(Sd)
Introduced themselves and explained the procedure to the patient.	0.001	0.196	1.86 (0.5)	1.88 (0.4)	1.91 (0.4)	2 (0.0)	1.95 (0.3)
Checked the accuracy and suitability of the medication and verified it was the correct patient (verbal confirmation is sufficient).	0.117	0.269	1.30 (0.9)	1.51 (0.9)	1.26 (0.9)	1.95 (0.2)	1.91 (0.4)
Ensured patient privacy (verbal confirmation is sufficient).	0.437	0.076	0.47 (0.9)	0.60 (0.8)	0.79 (0.9)	1.91 (0.3)	1.86 (0.5)
Syringe Preparation (Checked the expiration date of the syringe and ensured the packaging was intact), stated that they prepared the medication by drawing it into the syringe.	0.125	0.054	1.05 (0.4)	1.12 (0.9)	1.21 (0.5)	1.63 (0.5)	1.56 (0.6)
Indicated the injection site visually (upper outer quadrant).	0.043	0.072	1.70 (0.6)	1.81 (0.5)	1.84 (0.5)	1.98 (0.2)	1.95 (0.2)
Cleaned the injection area with an antiseptic cotton swab, wiping outward in a circular motion from the center, covering a radius of approximately 5 cm.*	0.049	-0.007	1.81 (0.4)	1.98 (0.2)	1.91 (0.3)	2 (0.0)	1.67 (0.7)
Placed a dry cotton swab between the 4th and 5th fingers of the assisting hand.	0.002	0.065	1.77 (0.6)	1.79 (0.6)	1.88 (0.3)	1.88 (0.3)	1.72 (0.5)
Stretched the skin at the injection site using the thumb and index finger of the assisting hand.	-0.005	0.177	1.77 (0.7)	1.79 (0.6)	1.84 (0.5)	1.86 (0.4)	1.98 (0.2)
Held the syringe with the active hand as if holding a pen.	0.001	0.289	1.88 (0.5)	1.93 (0.3)	1.88 (0.4)	2 (0.0)	1.98 (0.2)
Inserted the needle perpendicular to the skin.	-0.007	0.071	1.91 (0.4)	1.93 (0.3)	1.91 (0.4)	1.98 (0.2)	1.88 (0.3)
Pulled back the syringe plunger slightly with the assisting hand to check for blood.	-0.005	0.077	1.81 (0.5)	1.84 (0.4)	1.95 (0.2)	1.86 (0.4)	1.72 (0.7)
Injected the medication into the muscle by pressing the syringe plunger with the assisting hand (no actual medication will be injected into the model)	0.033	0.184	1.74 (0.5)	1.77 (0.5)	1.88 (0.3)	2 (0.0)	1.65 (0.8)
Removed the needle at the same angle and speed as insertion while applying light pressure to the injection site with a cotton swab using the assisting hand.	0.005	0.030	1.98 (0.2)	1.98 (0.2)	1.95 (0.2)	2 (0.0)	1.88 (0.4)
Disposed of the syringe and other waste materials in appropriate disposal containers (without recapping the needle).	0.058	0.056	1.60 (0.7)	1.86 (0.4)	1.91 (0.4)	1.98 (0.2)	1.72 (0.6)
Informed the patient that the procedure was complete and/or said, "Get well soon."	0.044	0.330	1.67 (0.8)	1.77 (0.6)	1.74 (0.7)	2 (0.0)	2 (0.0)

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, **α**: Krippendorff's, **κ**: Fleiss

Table 2 Comparison of evaluators

Analysis	Evaluator	*N*	M(Sd)	Test	*p*	Significant Difference
First Analysis	H1	43	24.33(3.2)	15.023*	< 0.001	H1 < AI1
	H2	43	25.56(3.4)			H1 < AI2
	RT	43	25.86(2.7)			H2 < AI1
	AI1	43	29.02(1.3)			RT < AI1
	AI2	43	27.44(4.1)			
Second Analysis	(1) Human Evaluating from Video	86	24.94(3.3)	24.665*	< 0.001	1 < 3
	(2) Real Time Human Evaluator	43	25.86(2.7)			2 < 3
	(3) Artificial Intelligence Evaluator	86	28.23(3.1)			
Third Analysis	(1) Human Evaluating	129	25.25(3.2)	-6.822**	< 0.001	1 < 2
	(2) Artificial Intelligence Evaluator	86	28.23(3.1)			

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, *ANOVA Test, **Independent Sample T-Test

who reviewed video recordings and the evaluations by ChatGPT-4o and Gemini Flash 1.5. The inter-rater reliability across these evaluators was measured using Krippendorff's Alpha and Fleiss' Kappa coefficients. Additionally, the arithmetic mean and standard deviation of the evaluations for the 9 criteria of the square knot skill, as assessed by evaluators for the 58 students, were analyzed. The results are presented in Table 3.

Krippendorff's Alpha values in Table 3 reveals that inter-rater reliability among the three human evaluators and two AI evaluators was not achieved across all 9 criteria for the square knot skill. The highest consistency was observed for the criterion "checked the tightness of the knot," while the lowest consistency was observed for "squeezed the string between the thumb and index finger of the other hand after crossing." Fleiss' Kappa values, calculated as an indicator of inter-rater consistency, showed the highest consistency for the criterion "squeezed the string between the thumb and index finger of the other hand after crossing," while the lowest consistency was found for "checked the tightness of the knot."

In nearly all criteria, human evaluators assigned lower scores compared to AI evaluators. Additionally, the standard deviation of scores provided by human evaluators was higher than that of AI evaluators, indicating that the scores given by human evaluators were more variable. The total scores for the 9 criteria of the square knot skill were calculated for each student, based on assessments by both human and AI evaluators. Subsequently, three types of comparisons were made: (1) A comparison analysis was conducted for the evaluations of each evaluator regarding the students' skills, (2) The assessments by the two human evaluators who reviewed video recordings were compared with those of the human evaluator who conducted the exam and the AI evaluators, (3) The human evaluators were grouped as one category, and the AI evaluators as another, and a comparison was made between the two groups. The results of these comparisons are presented in Table 4.

When Table 4 was examined, it was observed that human evaluators provided significantly lower scores compared to AI evaluators ($p < .05$).

Basic life support skill assessment

The assessment results for 47 students, evaluated by an examiner during the exam for basic life support (BLS) skills, were analyzed alongside the results of two human evaluators who reviewed video recordings and the assessments conducted by ChatGPT-4o and Gemini Flash 1.5. The inter-rater reliability across these evaluators was measured using Krippendorff's Alpha and Fleiss' Kappa coefficients. Additionally, the arithmetic mean and standard deviation of the evaluations for the 10 criteria of basic life support, as assessed by evaluators for the 47 students, were analyzed. The results are presented in Table 5.

When the Krippendorff's Alpha values in Table 5 were examined, it was found that inter-rater reliability was not achieved across all 10 criteria for basic life skills by the three human evaluators and two AI evaluators. The

Table 3 Mean, standard deviation values, and Inter-Rater reliability level of evaluators

Square Knot Criteria	Inter-rater Reliability		Evaluator				
			H1	H2	RT	AI1	AI2
	α	κ	M(Sd)	M(Sd)	M(Sd)	M(Sd)	M(Sd)
While holding both ends of the string in the palms of both hands, used the thumb of the non-dominant hand to pull the string in the other hand to create a cross.	0.170	0.093	1.14 (0.9)	1.28 (0.7)	1.60 (0.7)	2 (0.0)	1.74 (0.4)
Positioned the string between the thumb and index finger of the other hand after crossing.	0.088	0.110	0.98 (1.0)	1.05 (0.9)	1.55 (0.9)	1.66 (0.5)	1.67 (0.5)
Passed the string through the crossed section.	0.119	0.087	1.62 (0.8)	1.17 (0.9)	1.53 (0.8)	1.98 (0.1)	1.64 (0.6)
Adjusted the string's ends to a neat position and tied a square knot.	0.169	0.047	1.69 (0.6)	1.07 (0.9)	1.66 (0.6)	1.98 (0.1)	1.67 (0.5)
For the reverse knot, took the string from the thumb side of the cross, creating a new cross.	0.174	0.043	0.79 (0.9)	0.72 (0.9)	1.26 (0.8)	1.60 (0.5)	1.64 (0.5)
Brought the thumb and index fingers together and rolled the cross inside the hands to the other side.	0.248	0.066	0.84 (0.9)	0.79 (0.9)	1.26 (0.8)	1.97 (0.2)	1.67 (0.5)
Squeezed the other string between the thumb and index fingers, passing it back through the cross to the front side.	0.212	0.081	1.14 (0.9)	0.76 (0.9)	1.22 (0.9)	1.93 (0.3)	1.66 (0.5)
Straightened the ends of the string and tightened the knot.	0.259	0.033	1.60 (0.7)	0.71 (0.9)	1.41 (0.8)	1.98 (0.1)	1.59 (0.5)
Checked the tightness of the knot.	0.350	0.002	0.74 (0.9)	0.78 (0.8)	0.95 (0.9)	2 (0.0)	1.76 (0.5)

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, **α**: Krippendorff's, **κ**: Fleiss

Table 4 Comparison of evaluators

Analysis	Evaluator	*N*	M(Sd)	Test	*p*	Significant Difference
First Analysis	H1	58	10.55(5.8)	27.268*	< 0.001	H1 < AI1
	H2	58	8.33(7)			H1 < AI2
	RT	58	12.45(5.9)			H2 < RT
	AI1	58	17.10(0.9)			H2 < AI1
	AI2	58	15.03(3.4)			H2 < AI2
						RT < AI1
Second Analysis	(1) Human Evaluating from Video	116	9.44(6.5)	47.968*	< 0.001	1 < 2
	(2) Real Time Human Evaluator	58	12.45(5.9)			1 < 3
	(3) Artificial Intelligence Evaluator	116	16.07(2.7)			2 < 3
Third Analysis	(1) Human Evaluating	174	10.44(6.4)	-8.913**	< 0.001	1 < 2
	(2) Artificial Intelligence Evaluator	116	16.07(2.7)			

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, *ANOVA Test, **Independent Sample T-Test

Table 5 Mean, standard deviation values, and Inter-Rater reliability level of evaluators

Basic Life Support Criteria	Inter-rater Reliability		Evaluator				
			H1	H2	RT	AI1	AI2
	α	κ	M(Sd)	M(Sd)	M(Sd)	M(Sd)	M(Sd)
Checked the safety of the environment, themselves, and the patient (verbally stating this is sufficient).	0.267	0.100	0.94 (0.9)	1.51 (0.5)	0.96 (0.9)	2 (0.0)	1.89 (0.4)
Gently touched the patient/injured person's shoulders and asked, "How are you? Are you okay?" (verbally stating this is sufficient).	0.042	0.038	1.87 (0.5)	1.96 (0.2)	1.94 (0.3)	1.79 (0.4)	1.98 (0.2)
If the patient is unconscious, called for help from the environment and gave the command "Call 112" to someone (verbally stating this is sufficient).	0.017	0.161	1.87 (0.5)	1.96 (0.2)	1.77 (0.5)	1.81 (0.4)	1.89 (0.4)
Checked the mouth, opened the airway using the "head-tilt, chin-lift" maneuver.	0.015	0.023	1.70 (0.6)	1.94 (0.3)	1.74 (0.5)	1.72 (0.5)	1.81 (0.5)
Assessed breathing for no more than 10 s using the "look, listen, feel" method and checked for pulse at the carotid artery.	0.051	0.024	1.60 (0.7)	1.96 (0.2)	1.57 (0.7)	1.87 (0.3)	1.68 (0.6)
Performed effective and correct chest compressions (correct hand position, correct compression point, correct depth, correct speed, and allowing chest recoil).	0.293	-0.014	1.55 (0.6)	2 (0.0)	1.34 (0.7)	1.66 (0.5)	1 (0.6)
After 30 chest compressions, effectively gave 2 rescue breaths with proper head-tilt and chin-lift position (closed the patient's nostrils while giving breaths).	0.151	0.071	1.87 (0.3)	1.98 (0.2)	1.49 (0.5)	1.55 (0.5)	1.68 (0.5)
Minimized interruptions in chest compressions.	0.046	0.042	1.53 (0.9)	2 (0.0)	1.64 (0.6)	1.83 (0.4)	1.70 (0.6)
Continued performing chest compressions and rescue breaths in a 30/2 ratio for two minutes (or stated that they should do so).	0.143	-0.033	1.68 (0.7)	2 (0.0)	1.77 (0.5)	1.96 (0.2)	1.36 (0.8)
Checked the patient's breathing and pulse every two minutes (verbally stating this is sufficient).	0.163	-0.003	0.57 (0.8)	1.45 (0.5)	1.28 (0.8)	1.62 (0.7)	1.30 (0.9)

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, **α**: Krippendorff's, **κ**: Fleiss

highest consistency was found in the criterion "performed effective and correct chest compressions (correct hand position, correct compression point, correct depth, correct speed, and allowed chest recoil)," while the lowest consistency was found in the criterion "checked the oral cavity and opened the airway using the 'head-tilt, chin-lift maneuver." When the Fleiss Kappa values, which measure inter-rater consistency, were examined, the highest consistency was found in the criterion "if unconscious, called for help from the environment and instructed someone to call '112' (verbal indication was sufficient)," while the lowest consistency was found in the criterion "continued chest compressions and rescue breathing in a 30/2 ratio for two minutes (or stated that it should be done)."

The total scores for each student on the 10 criteria of basic life skills were calculated for both human and AI evaluators. Following this, three different comparisons were made. First, the evaluations of each evaluator were compared through a comparative analysis of the students' skills. Second, the two human evaluators who assessed via video, the human evaluator who assessed during the exam, and the AI evaluators were compared. Third, human evaluators were grouped as one group, and AI evaluators were grouped as another for comparison. The results are provided in Table 6.

Table 6 shows that human evaluators scored significantly lower than compared AI evaluators ($p < .05$).

Table 6 Comparison of evaluators

Analysis	Evaluator	*N*	M(Sd)	Test	*p*	Significant Difference
First Analysis	H1	47	15.19(2.6)	20.052	0.001	H1 < H2 H1 < AI1 H2 > RT H2 > AI2 RT < AI1 AI1 > AI2
	H2	47	18.74(1.0)			
	RT	47	15.49(2.9)			
	AI1	47	17.81(0.9)			
	AI2	47	16.30(3.2)			
Second Analysis	(1) Human Evaluating from Video	94	16.97(2.6)	6.312	0.002	1 > 2 2 < 3
	(2) Real Time Human Evaluator	47	15.49(2.9)			
	(3) Artificial Intelligence Evaluator	94	17.05(2.5)			
Third Analysis	(1) Human Evaluating	141	16.48(2.8)	-1.620	0.107	None
	(2) Artificial Intelligence Evaluator	94	17.05(2.5)			

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, *ANOVA Test, **Independent Sample T-Test

Urinary catheter placement skill assessment

The evaluation results of 48 students who participated in the urinary catheter placement skill exam were analyzed for consistency between the results of the evaluator during the exam, the two human evaluators who assessed the video, and the evaluations conducted by ChatGPT-4o and Gemini Flash 1.5. The consistency was examined using Krippendorff's Alpha and Fleiss Kappa coefficients. Additionally, the arithmetic mean and standard deviation of the assessments made by the evaluators for the 15 criteria of urinary catheter placement skill for the 48 participants were also analyzed. The results are presented in Table 7.

When examining Krippendorff's Alpha values in Table 7, it was found that there was no inter-rater reliability for all 15 criteria for urinary catheterization skill between three human raters and two AI raters. The highest consistency was observed in the criterion "suspended the urine bag below the bladder," while the lowest consistency was found in the criterion "attached the urine bag (it could have been attached earlier)." When examining the Fleiss Kappa values as indicators of inter-rater consistency, the highest consistency was observed in the criterion "attached the urine bag (it could have been attached earlier)," while the lowest consistency was observed in the criterion "suspended the urine bag below the bladder." The lowest scores were given by the first human rater evaluating from the video. The variance of the scores given by this same rater was also high. Among the AI raters, the evaluation with high variance was attributed to Gemini Flash 1.5. The highest scores were given by the second human rater evaluating from the video and the human rater evaluating in real-time during the exam.

The total scores for each student were obtained based on the evaluations of human and AI raters on the 15 criteria of urinary catheterization skill for 48 students. Following this, three different comparisons were made. First, the evaluations made by each rater on the students' skills were compared using comparison analysis. Secondly, the evaluations of two human raters evaluating from the video, the human rater evaluating during the exam, and the AI raters were compared. Third, human raters were grouped into one group, and AI raters were grouped into another, and a comparison was made between these groups. The results are provided in Table 8.

Table 8 revealed that the second human rater evaluating from the video and the real-time rater gave significantly higher scores ($p < .05$) than the second human rater evaluating from the video and the AI raters.

In Krippendorff's Alpha and Fleiss Kappa inter-rater reliability analyses, it was observed that there was significant variation in the scores among human raters. Due to these findings, the average of the three human raters' evaluations was accepted as a "gold standard," representing the general trend of human expertise by minimizing individual expert differences [23]. This was then compared separately with the evaluations of AI models, and Bland-Altman analyses were conducted to assess agreement between AI models and the average scores of human evaluators across the four clinical skills. In all comparisons, both ChatGPT-4o and Gemini Flash 1.5 models demonstrated a tendency to assign higher total scores than human raters, with negative bias values indicating this upward scoring trend. The total score bias was greatest in the square knot tying skill (ChatGPT-4o: -6.66; Gemini Flash 1.5: -4.59), followed by intramuscular injection (ChatGPT-4o: -3.78; Gemini Flash 1.5: -2.19), reflecting systematic scoring differences between AI and human evaluations.

Agreement varied by perception type. In general, visual criteria (e.g., injection) yielded closer alignment between human and AI scores, often with bias values near zero. In contrast, auditory criteria (e.g., obtaining verbal consent, giving instructions) resulted in larger deviations and lower agreement, especially with ChatGPT-4o (e.g., bias of -1.29 for ensuring privacy verbally during injection).

When evaluating the limits of agreement (LOA), most data points fell within acceptable boundaries (± 1.96 SD),

Table 7 Mean, standard deviation values, and Inter-Rater reliability level of evaluators

Urinary Catheter Placement Criteria	Inter-rater Reliability		Evaluator				
			H1	H2	RT	AI1	AI2
	α	κ	M(Sd)	M(Sd)	M(Sd)	M(Sd)	M(Sd)
Introduced themselves to the patient.	0.139	-0.028	1.79 (0.6)	2 (0.0)	1.98 (0.1)	1.65 (0.6)	2 (0.0)
Explained the procedure to be performed and obtained consent for pelvic examination (informed consent obtained).	0.156	-0.041	1.48 (0.7)	1.98 (0.1)	2 (0.0)	1.81 (0.4)	1.81 (0.5)
Washed and dried their hands (verbal explanation was sufficient).	0.073	-0.001	1.48 (0.9)	1.92 (0.3)	2 (0.0)	1.85 (0.4)	1.90 (0.4)
Wore sterile gloves (saying it was sufficient).	0.102	-0.011	1.60 (0.7)	1.96 (0.2)	2 (0.0)	1.92 (0.3)	1.90 (0.4)
Indicated the injection site visually (upper outer quadrant).	0.297	-0.108	1.35 (0.5)	2 (0.0)	1.98 (0.1)	1.73 (0.5)	1.85 (0.4)
Cleaned the penis three times with gauze soaked in antiseptic solution, starting from the external urethral orifice in expanding circular motions.	0.103	-0.019	1.94 (0.3)	2 (0.0)	2 (0.0)	1.85 (0.4)	1.54 (0.8)
Applied lubricant gel to the tip of the chosen Foley catheter.	0.026	-0.018	1.92 (0.4)	1.96 (0.2)	1.92 (0.4)	1.85 (0.4)	1.79 (0.4)
With one hand, directed the tip of the catheter while holding the penis with the other hand, gently advancing the catheter tip into the urethra.	0.311	-0.032	0.83 (0.9)	1.98 (0.1)	2 (0.0)	1.85 (0.4)	1.71 (0.5)
Held the syringe with the active hand as if holding a pen.	0.146	-0.034	1.27 (0.9)	2 (0.0)	1.85 (0.4)	1.88 (0.3)	1.96 (0.2)
When the catheter tip reached the perineal level, lowered the patient's penis and aligned it parallel to the body's long axis.	0.018	-0.009	1.75 (0.7)	1.96 (0.2)	1.92 (0.4)	1.92 (0.3)	1.71 (0.6)
Continued to advance the catheter upward, parallel to the body's long axis.	0.124	-0.041	1.33 (0.9)	1.98 (0.1)	1.90 (0.4)	1.85 (0.4)	1.46 (0.7)
Observed the urine flow from the expected tip of the catheter / noted the urine output.	-0.002	0.061	1.85 (0.5)	1.96 (0.2)	1.94 (0.3)	1.85 (0.4)	1.85 (0.5)
After urine flow started, advanced the catheter 2–3 cm further.	0.122	-0.010	1.83 (0.6)	1.98 (0.1)	2 (0.0)	1.79 (0.4)	1.48 (0.8)
Attached the urine bag (it could have been attached earlier).	0.206	-0.043	1.13 (1.0)	1.98 (0.1)	1.88 (0.3)	1.92 (0.3)	1.40 (0.7)
Administered saline to the balloon's pathway using a syringe	0.339	-0.055	0.69 (0.9)	1.96 (0.2)	1.85 (0.4)	1.94 (0.3)	1.33 (0.8)
Slowly pulled the catheter back and, after feeling the balloon rest on the bladder neck, advanced it 1–2 cm.	0.139	-0.028	1.79 (0.6)	2 (0.0)	1.98 (0.1)	1.65 (0.6)	2 (0.0)
Suspended the urine bag below the bladder, ensuring proper positioning	0.156	-0.041	1.48 (0.7)	1.98 (0.1)	2 (0.0)	1.81 (0.4)	1.81 (0.5)

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, **α**: Krippendorff's, **κ**: Fleiss

Table 8 Comparison of evaluators

Analysis	Evaluator	*N*	M(Sd)	Test	*p*	Significant Difference
First Analysis	H1	48	22.25(4.2)	54.152	0.001	H1 < H2, H1 < RT
	H2	48	29.60(1.1)			H1 < AI1, H1 < AI2
	RT	48	29.21(1.8)			H2 > AI1, H2 > AI2
	AI1	48	27.67(1.8)			RT > AI2, AI1 > AI2
	AI2	48	25.69(3.9)			
Second Analysis	(1) Human Evaluating from Video	96	25.93(4.8)	12.680	0.001	1 < 2
	(2) Real Time Human Evaluator	48	29.21(1.8)			2 > 3
	(3) Artificial Intelligence Evaluator	96	26.68(3.2)			
Third Analysis	(1) Human Evaluating	144	27.02(4.3)	0.668	0.505	None
	(2) Artificial Intelligence Evaluator	96	26.68(3.2)			

H1: Human Evaluating from Video 1, **H2**: Human Evaluating from Video 2, **RT**: Real Time Human Evaluator, **AI1**: ChatGPT, **AI2**: Gemini Flash, **M**: Mean, **Sd**: Standard Deviation, *ANOVA Test, **Independent Sample T-Test

yet some outliers were noted, especially in tasks requiring fine motor skills or precise verbal responses. Gemini Flash 1.5 generally demonstrated a narrower LOA range than ChatGPT-4o, indicating slightly more stable agreement, particularly in visually oriented tasks. However, both models exhibited greater variability and bias when evaluating complex or nuanced actions that required auditory interpretation or subtle visual cues.

These analyses highlight that while AI models can replicate human scoring trends for structured, visual tasks, their performance declines in skills involving auditory input or subjective interpretation. Complete statistical outputs, visual Bland-Altman plots, tables and extra explanations are provided in the Supplementary File 1.

Discussion

Our findings provide significant insights into the potential role of AI in medical education assessments, particularly in OSCE evaluations. The study revealed that AI-based evaluators, specifically ChatGPT-4o and Gemini Flash 1.5, consistently assigned higher scores compared to human evaluators. The agreement between AI and human evaluators varied across clinical skills, with higher consistency observed in tasks requiring visual perception compared to those involving auditory, visual or auditory-visual inputs. While AI models show potential for objective and consistent assessments, they may overestimate student performance in certain skills. The higher scores provided by AI could reflect a more lenient interpretation of evaluation criteria or limitations in detecting nuanced errors. For instance, tasks requiring auditory inputs, such as ensuring patient privacy with verbal confirmation, exhibited the lowest inter-rater reliability, indicating that current AI systems may struggle with evaluating context-specific auditory cues.

The highest agreement levels between AI and human evaluators were observed in straightforward visual tasks, such as identifying injection sites and inserting needles correctly. This finding underscores the strength of AI in visually dominant skills but highlights its limitations in evaluating procedural steps that rely heavily on auditory feedback or interpersonal communication. A possible explanation for this discrepancy is the lack of auditory-specific training data or the absence of real-time contextual understanding in AI models. Future iterations of AI systems should aim to integrate multimodal inputs more effectively to bridge this gap. It is also worth noting that this study was conducted with untrained AI models; better results might be achieved with trained models. Thus, similar studies with trained models should be pursued to evaluate their potential in clinical skill assessments.

Above-mentioned findings highlight the need for further refinement in AI-based OSCE evaluation tools, particularly in improving their capacity to assess complex, multimodal clinical skills accurately. Additionally, further research is needed to explore the potential biases and limitations of AI assessments, as well as their long-term impact on medical education outcomes. We believe these findings will pave the way for integrating AI into medical education assessments, ensuring reliability and fairness. Our findings revealed that AI systems demonstrated higher inter-rater reliability in visual tasks, aligning with previous studies that emphasize AI's strength in image-based assessments [11, 24]. However, its performance in auditory and multimodal tasks highlighted limitations, consistent with findings from Chan & Tsi [1] and Yamamoto et al. [25].

Substantial agreement between AI and human evaluators in visually dominant criteria supports prior literature suggesting that AI excels in pattern recognition and objective measurements [13, 15]. For instance, AI's ability to consistently assess suturing techniques aligns with Fazlollahi et al. [26], who demonstrated AI's accuracy in evaluating procedural skills. However, in the specific context of knot-tying skills, particularly square knot techniques, AI models generated more variations and encountered difficulties in assessing fine motor movements and visual details. This tendency for AI models to consistently assign higher scores in tasks involving fine motor skills suggests potential limitations in evaluating these skills accurately and raises concerns about possible hallucinatory interpretations of subtle hand movements or visual cues. These findings underscore the need for hybrid evaluation models that leverage AI's strengths while addressing its limitations through human oversight [27]. Such integration could reduce rater bias and provide objective feedback, a key concern in traditional OSCEs [28]. Furthermore, the hybrid use of expert-based evaluators and AI models in clinical skills assessments could offer a powerful feedback mechanism, enabling their application as coaching tools or peer educators in medical training.

Our results align with Sallam et al. [29], who reported that AI systems outperformed humans in standardized assessments but struggled with contextual and interpretive tasks. Similarly, Chan & Lee [3] highlighted AI's limitations in dynamic clinical interactions. These studies collectively suggest that while AI can enhance efficiency and objectivity, its application in complex, interpretive scenarios requires further refinement. Moreover, our study contributes to the growing body of literature advocating for AI as a supplementary tool in medical education [30]. By identifying specific areas where AI underperforms, future research can focus on improving algorithmic designs to better interpret multimodal data [31].

This study acknowledges several limitations. The sample size, though sufficient for initial analysis, may

not capture the full variability of clinical performance across diverse populations, potentially affecting the generalizability of the findings. Additionally, the reliance on a single AI system may have limited the scope of the results, as different platforms could yield varying levels of accuracy and consistency [32]. These limitations suggest that certain nuances in clinical performance might have been overlooked, and future research should address this by exploring multi-institutional datasets and evaluating different AI models to ensure broader applicability [33]. Furthermore, while our findings demonstrate AI's potential, ethical concerns regarding data privacy and algorithmic bias remain unresolved. These issues require multidisciplinary collaboration to ensure responsible implementation [34]. In addition, it is important to recognize that AI systems may introduce other forms of bias. These can stem from limitations in training data, algorithmic assumptions, or insufficient sensitivity to cultural, contextual, or linguistic diversity. For instance, an AI model may misinterpret performance nuances in students who speak with different accents, use alternative communication styles, or follow slightly varied but acceptable procedural steps. As AI becomes more integrated into assessment processes, careful validation and auditing are needed to identify and mitigate such risks.

The study's findings have several implications for medical education. First, incorporating AI into OSCEs could standardize evaluations and reduce rater inconsistencies, ultimately enhancing the reliability of assessments [35, 36]. Second, AI-driven feedback systems could provide students with actionable insights, promoting targeted skill development [32].

Conclusion

This study provides a comprehensive analysis of the potential of AI systems in medical education, particularly in OSCE evaluations. Our study revealed that AI-based evaluators, particularly ChatGPT-4o and Gemini Flash 1.5, tend to assign higher and more consistent scores than human evaluators in OSCE assessments across four different clinical skills. Agreement between human and AI assessments was stronger for visually based tasks, while discrepancies were more pronounced in steps requiring auditory interpretation. Although current AI models show potential as supplementary evaluators in clinical skills assessment, further refinement is needed to ensure consistency with human standards, particularly in evaluating communication- and audio-dependent steps.

Future research should explore integrating AI with advanced multimodal learning systems to address its current limitations. Additionally, longitudinal studies are needed to assess AI's impact on long-term skill retention and clinical outcomes [36, 37]. Finally, developing ethical guidelines for AI implementation in medical education is critical to maintaining trust and ensuring equitable access [33]. In addition, further studies should focus on developing and testing trained AI models to address the identified limitations and explore their application in diverse clinical contexts. Additionally, longitudinal studies are required to evaluate the impact of AI-driven assessments on long-term clinical competency and professional growth. By leveraging the strengths of AI while addressing its weaknesses, medical education can move towards more standardized, efficient, and fair evaluation systems, ultimately improving the training and readiness of future healthcare professionals.

Abbreviations

OSCE	Objective Structured Clinical Examinations
AI	Artificial intelligence
LLM	Large Language Model
M-LLM	Multimodal Large Language Models
OCR	Optical character recognition
IM	Intramuscular
LOA	Limits of agreement
BLS	Basic life support

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-025-07241-4.

Supplementary Material 1

Supplementary Material 2

Acknowledgements

The authors would like to thank all students that participated in the study.

Author contributions

MOY, MT and IU collected and ÇT, MOY and GK analyzed the data, and each author took equal responsibility for writing, revising and editing of the manuscript. All authors read and approved the final manuscript.

Funding

The authors was not receive any financial support for the research, authorship, or publication of this study.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study was conducted with the approval of the Çanakkale Onsekiz Mart University Non-Interventional Clinical Research Ethics Committee (Date of Approval: 03.06.2024/No: 2024/06–08). Our research was conducted in accordance with the Declaration of Helsinki, and informed consent was obtained from all individual participants included in the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 10 January 2025 / Accepted: 25 April 2025
Published online: 01 May 2025

References

1. Shorey S, Chan V, Rajendran P, Ang E. Learning styles, preferences and needs of generation Z healthcare students: scoping review. Nurse Educ Pract. 2021;57:103247.
2. Chan CKY, Tsi LH. The AI Revolution in Education: Will AI Replace or Assist Teachers in Higher Education? arXiv preprint arXiv:2305.01185. 2023.
3. Chan CKY, Lee KKW. The AI generation Gap: are gen Z students more interested in adopting generative AI such as ChatGPT in teaching and learning than their gen X and millennial generation teachers? Smart Learn Environ. 2023;10:60. https://doi.org/10.1186/s40561-023-00269-3.
4. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J. 1975;1(5955):447–51.
5. Patrício MF, Julião M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach. 2013;35(6):503–14. https://doi.org/10.3109/0142159X.2013.774330.
6. Azcuenga J, Valls M, Martinez-Carretero JM. Cost analysis of three clinical skills assessment (CSA) projects. Protecting the human dimension. Evol Assess. 1998;8th Ottawa Conference(1998):Philadelphia.
7. Carpenter JL. Cost analysis of objective structured clinical examinations. Acad Med. 1995;70(9):828–33.
8. Poenaru D, Morales D, Richards A, O'Connor HM. Running an objective structured clinical examination on a shoestring budget. Am J Surg. 1997;173(6):538–41.
9. Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, Berard M. Guidelines for estimating the real cost of an objective structured clinical examination. Acad Med. 1993;68(7):513–7.
10. Cusimano MD. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). Acad Med. 1994;69:571–6.
11. Soong TK, Ho CM. Artificial Intelligence in Medical OSCEs: Reflections and Future Developments. Advances in Medical Education and Practice. 2021; 12: 167–173. https://doi.org/10.2147/AMEP.S287926
12. Xie J, Chen Z, Zhang R, Wan X, Li G. Large Multimodal Agents: A Survey. arXiv preprint arXiv:2402.15116. 2024.
13. Wu S, Fei H, Qu L, Ji W, Chua TS. Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519. 2023.
14. Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549. 2023.
15. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Seifman MA. Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. ANZ J Surg. 2024;94(1–2):68–77.
16. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.
17. Fleiss JL. Measuring nominal scale aagreement among many raters. Psychol Bull. 1971;76(5):378–82. https://doi.org/10.1037/h0031619.
18. Landis R, Koch G. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74. https://doi.org/10.2307/2529310.
19. Krippendorff K. Content analysis: An introduction to its methodology (2nd Edition). ISBN 0-7619-1545-1. USA: Sage Publications; 2004.
20. Krippendorff K. Computing Krippendorff's Alpha-Reliability. Retrieved from http://repository.upenn.edu/asc_papers/43. 2011.
21. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;327(8476):307–10.
22. Giavarina D. Understanding Bland-Altman analysis. Biochemia Med. 2015;25(2):141–51.
23. Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters (4th ed.). Advanced Analytics, LLC. 2014.
24. Bicknell BT, Butler D, Whalen S, Ricks J, Dixon CJ, Clark AB, Spaedy O, Skelton A, Edupuganti N, Dzubinski L, Tate H, Dyess G, Lindeman B. Lehmann, LS. ChatGPT-4 omni performance in USMLE disciplines and clinical skills: comparative analysis. JMIR Med Educ. 2024; 10(1), e63430.
25. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, Ino H. Enhancing medical interview skills through AI-Simulated patient interactions: nonrandomized controlled trial. JMIR Med Educ. 2024; 10(1), e58753.
26. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, Langleben I, Ledwos N, Sabbagh AJ, Bajunaid K, Harley JM, Del Maestro RF. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. JAMA Netw Open. 2022;5(2):e2149008–2149008.
27. Johnsson V, Søndergaard MB, Kulasegaram K, Sundberg K, Tiblad E, Herling L, Petersen OB, Tolsgaard MG. Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Med Educ. 2024;58(1):105–17.
28. Mortlock R, Lucas C. Generative artificial intelligence (Gen-AI) in pharmacy education: utilization and implications for academic integrity: A scoping review. Exploratory Res Clin Social Pharm. 2024;15:100481.
29. Sallam M, Al-Salahat K, Eid H, Egger J, Puladi B. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, bard, ChatGPT-3.5 and humans in clinical chemistry Multiple-Choice questions. Adv Med Educ Pract. 2024;15:857–71.
30. Maurya RK. Using AI based chatbot ChatGPT for practicing counseling skills through role-play. J Creativity Mental Health. 2024;19(4):513–28.
31. Tolentino R, Baradaran A, Gore G, Pluye P, Abbasgholizadeh-Rahimi S. Curriculum frameworks and educational programs in AI for medical students, residents, and practicing physicians: scoping review. JMIR Med Educ. 2024; 10(1), e54793.
32. Jamal A, Solaiman M, Alhasan K, Temsah MH, Sayed G. Integrating ChatGPT in medical education: adapting curricula to cultivate competent physicians for the AI era. Cureus. 2023;15(8):e43036. https://doi.org/10.7759/cureus.43036.
33. Knopp MI, Warm EJ, Weber D, Kelleher M, Kinnear B, Schumacher DJ, Santen SA, Mendonça E, Turner. L. AI-enabled medical education: threads of change, promising futures, and risky realities across four potential future worlds. JMIR Med Educ. 2023; 25(9), e50373.
34. Turner L, Hashimoto DA, Vasisht S, Schaye V. Demystifying AI: current state and future role in medical education assessment. Acad Med. 2023;99:42–7.
35. Ghorashi N, Ismail A, Ghosh P, Sidawy A, Javan R. AI-powered chatbots in medical education: potential applications and implications. Cureus, 2023; 15(8), e43271.
36. 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;11:23821205241263475.
37. Sharma P, Thapa K, Thapa D, Dhakal P, Upadhaya MD, Adhikari S, Khanal SR. Performance of ChatGPT on USMLE: Unlocking the potential of large language models for AI-assisted medical education. arXiv preprint arXiv:2307.00112. 2023.

Publisher's note