Artículo de revisión

MECP | MEDICAL EDUCATION & CLINICAL PRACTICE

El papel emergente de la inteligencia artificial en los exámenes clínicos objetivos estructurados

The emerging role of artificial intelligence in objective structured clinical examinations

Nelson Luis Cahuapaza-Gutierrez[1,2,*], Cielo Cinthya Calderon-Hernandez[1,2]

[1] Universidad Científica del Sur, Lima, Perú.
[2] Research Department, N y C – Center of Research and Medical Excellence (CRME), Lima, Perú.

RESUMEN

La evaluación de las competencias clínicas es fundamental en la formación médica, y el Examen Clínico Objetivo Estructurado (ECOE) constituye una herramienta ampliamente utilizada para este propósito. El desarrollo de la inteligencia artificial (IA) ha generado interés en su aplicación en los diferentes componentes del ECOE. En este contexto, se revisó la literatura reciente para sintetizar la evidencia disponible sobre sus aplicaciones. Los hallazgos sugieren que los modelos de IA, particularmente los grandes modelos de lenguaje podrían apoyar la enseñanza, el aprendizaje y la evaluación mediante la generación y resolución de estaciones, la calificación de evaluaciones, la provisión de retroalimentación y la mejora de habilidades clínicas para el ECOE. Sin embargo, presentan limitaciones para evaluar aspectos complejos de la comunicación clínica, como la comunicación no verbal, la empatía y el juicio clínico. Por ello, su uso debe considerarse complementario a la experiencia docente y clínica.

Palabras clave: Inteligencia Artificial; Examen Físico; Educación Médica (fuente: DeCS Bireme).

ABSTRACT

The assessment of clinical competencies is fundamental to medical education, and the Objective Structured Clinical Examination (OSCE) is widely used for this purpose. Advances in artificial intelligence (AI) have generated growing interest in its application across different components of the OSCE. In this context, we reviewed the recent literature to synthesize the available evidence on its applications. The findings suggest that AI models, particularly large language models (LLMs), could support OSCE-related teaching, learning, and assessment by generating and solving stations, scoring performance, providing feedback, and enhancing clinical skills. However, these models have limitations in assessing complex aspects of clinical communication, including nonverbal communication, empathy, and clinical judgment. Therefore, AI should be considered a complement to, rather than a replacement for, faculty and clinical expertise.

Keywords: Artificial Intelligence; Physical Examination; Medical Education (source: MeSH NLM).

INTRODUCCIÓN

La evaluación de las competencias clínicas constituye uno de los pilares fundamentales de la formación de los profesionales de la salud. Entre las estrategias desarrolladas para este propósito, el Examen Clínico Objetivo Estructurado (Objective Structured Clinical Examination, OSCE), conocido en muchos países de habla hispana como ECOE, se ha consolidado como una herramienta ampliamente utilizada para evaluar habilidades clínicas, comunicación médico-paciente, razonamiento diagnóstico y toma de decisiones en entornos simulados y estandarizados [1]. Su diseño basado en estaciones, en las que se recrean escenarios clínicos específicos, permite estructurar la evaluación de manera objetiva y reproducible. Sin embargo, su implementación requiere una considerable inversión de tiempo, recursos humanos y logística, particularmente para el diseño de escenarios, capacitación de evaluadores, calificación y provisión de retroalimentación [2-4].

Paralelamente, los avances en inteligencia artificial (IA), impulsados por el desarrollo de redes neuronales profundas (*deep learning*), la disponibilidad de grandes volúmenes de datos y el incremento de la capacidad computacional, han ampliado las posibilidades de aplicación de estas tecnologías en educación médica [5]. En particular, los grandes modelos de lenguaje (LLM; *Large Language Models*)

Historial del artículo
Recibido: 15 de Junio de 2026
Revisado por pares
Aceptado: 19 de Agosto de 2026
En línea: 28 de Agosto de 2026

Cómo citar el artículo
Cahuapaza-Gutierrez NL, Calderon-Hernandez CC. El papel emergente de la inteligencia artificial en los exámenes clínicos objetivos estructurados. Med Educ Clin Pract. 2026;1(3):e032. doi: 10.21142/mecp.2026.1.3.e032.

***Correspondencia**
Nelson Luis Cahuapaza-Gutierrez
100065659@cientifica.edu.pe

han mostrado capacidades para procesar lenguaje clínico, generar contenido educativo e interactuar con usuarios mediante respuestas adaptativas, mientras que otros sistemas de IA permiten analizar información multimodal y apoyar diferentes procesos de evaluación y aprendizaje [6-9]. Como consecuencia, la IA se está incorporando progresivamente a la educación médica y ha despertado interés en su aplicación a los diferentes componentes de los ECOE [10].

La literatura reciente describe diversas aplicaciones de la IA en este contexto. En la generación de estaciones, los LLM pueden apoyar la elaboración de escenarios clínicos, guiones para pacientes simulados, listas de verificación y preguntas alineadas con objetivos específicos de aprendizaje, lo que podría reducir el tiempo requerido para el desarrollo de estos materiales y favorecer una mayor estandarización [11]. Asimismo, algunos sistemas basados en IA y procesamiento de lenguaje natural se han utilizado para analizar respuestas verbales, interacciones clínicas y desempeño de los estudiantes, así como para apoyar procesos de calificación [12,13]. Estas aplicaciones podrían contribuir a reducir la variabilidad entre evaluadores y facilitar procedimientos de evaluación más consistentes. Otra área de interés corresponde a la retroalimentación. Los sistemas basados en IA pueden generar comentarios personalizados y oportunos a partir del desempeño del estudiante, identificando fortalezas, áreas de mejora y posibles estrategias de aprendizaje [14-16]. Además, la IA se ha utilizado como herramienta de entrenamiento mediante pacientes virtuales y simulaciones interactivas que permiten practicar repetidamente habilidades como la anamnesis, el razonamiento diagnóstico y la comunicación clínica [12,17]. Estas aplicaciones pueden ofrecer oportunidades de práctica flexibles y adaptadas al ritmo del estudiante, complementando las posibilidades de entrenamiento disponibles mediante los ECOE tradicionales [18-20].

A pesar del creciente número de aplicaciones descritas, persisten incertidumbres sobre su precisión, confiabilidad, validez y capacidad para reproducir adecuadamente aspectos complejos de la evaluación clínica [21,22]. La comunicación no verbal, la empatía contextual, la interacción con el paciente y el juicio clínico continúan siendo limitaciones de la evaluación automatizada. Asimismo, los resultados sobre su desempeño frente a evaluadores humanos e impacto educativo no son uniformes. La evidencia sobre su aplicación en los ECOE permanece fragmentada entre la generación y resolución de estaciones, la calificación, la retroalimentación y el fortalecimiento de habilidades clínicas, dificultando valorar integralmente sus beneficios, limitaciones y aplicaciones.

En este contexto, el objetivo de esta revisión de la literatura fue sintetizar las aplicaciones de la inteligencia artificial en los ECOE, incluyendo la generación y resolución de estaciones, la calificación, la provisión de retroalimentación y la mejora de habilidades clínicas, así como sintetizar los principales resultados y limitaciones reportados de los estudios.

METODOLOGÍA

Se realizó una revisión de la literatura con el objetivo de sintetizar la evidencia disponible sobre las aplicaciones actuales de la inteligencia artificial en el contexto de los ECOEs. La revisión muestra el uso de modelos de IA en diferentes etapas del ECOE, incluyendo generación y resolución de estaciones, la calificación de evaluaciones, la provisión de retroalimentación y la mejora de habilidades clínicas para el ECOE. La búsqueda bibliográfica se realizó en las bases de datos PubMed y Scopus desde su creación hasta el 11 de junio de 2026. Además, se recopilaron ejemplos representativos de los *prompts* utilizados con los diferentes modelos de IA, los cuales se presentan en el Material Suplementario. Los hallazgos se sintetizaron mediante un enfoque narrativo y temático con el fin de proporcionar una visión integral del estado actual de la evidencia.

Los dominios de análisis se definieron a partir de la literatura previa sobre evaluación mediante el ECOE, simulación clínica y aplicaciones de IA en educación médica. Considerando la heterogeneidad de las aplicaciones descritas en la literatura, se establecieron cinco dominios temáticos para organizar y sintetizar la evidencia: (I) generación de estaciones, (II) resolución de estaciones, (III) calificación de estaciones, (IV) retroalimentación y (V) habilidades clínicas. I) La generación de estaciones: comprende el desarrollo de escenarios clínicos simulados, controlados y reproducibles en los que los estudiantes interactúan con participantes estandarizados y son evaluados mediante criterios predefinidos. Los LLM han sido explorados para simular pacientes estandarizados, examinadores y estudiantes, así como para apoyar el desarrollo de estaciones de ECOE [23]. II) La resolución de estaciones: se refiere al uso de sistemas de inteligencia artificial para generar respuestas precisas, sofisticadas y contextualmente relevantes durante la resolución de escenarios clínicos simulados, en función de las instrucciones y demandas planteadas en una estación de ECOE [24]. III) La calificación de estaciones: se define como el uso de sistemas de IA para asignar, estimar la puntuación del desempeño de los estudiantes a partir de grabaciones de video. Para cada habilidad evaluada, se utilizan listas de verificación específicas [25]. IV) Retroalimentación: es un mecanismo

regulador mediante el cual la información derivada del desempeño se utiliza para modificar y mejorar acciones futuras. Para ser efectiva, debe ser específica, oportuna y práctica, y abordar tanto el contenido como el proceso del razonamiento clínico. Los avances recientes en los LLM permiten analizar de forma automatizada las interacciones entre estudiantes y pacientes virtuales, proporcionando retroalimentación escalable sobre la anamnesis y el razonamiento diagnóstico [12]. V) Habilidades clínicas: son competencias necesarias para la atención del paciente, que incluyen la entrevista médica, la comunicación clínica y la recopilación sistemática de información clínica, evaluadas mediante escenarios estandarizados y orientadas a preparar al estudiante para la práctica clínica [26].

RESULTADOS

Modelos de inteligencia artificial para responder estaciones del Examen Clínico Objetivo Estructurado

La evaluación clínica objetiva estructurada es fundamental para la formación y certificación de competencias clínicas, pero su diseño, implementación y calificación demandan recursos humanos, tiempo y estandarización. En este contexto, la IA, particularmente los LLM, representa una nueva frontera en la educación médica. Estudios previos han documentado su aplicación en exámenes médicos exigentes, como el Examen Nacional de Medicina en Perú (ENAM), el Examen de Acceso a la Residencia Médica en España (MIR) y el Examen de Licencia Médica de los Estados Unidos (USMLE, pasos 1 y 2), mostrando un rendimiento consistentemente alto y la superación de los umbrales de aprobación [27]. Sin embargo, su capacidad para resolver estaciones del ECOE permanece poco explorada. Su naturaleza multicomponente y orientada a competencias plantea desafíos diferentes de los exámenes convencionales de opción múltiple (Tabla 1).

Sarah Li *et al.* investigaron la capacidad de ChatGPT para interactuar y completar de forma autónoma un ECOE simulado [24]. Sus resultados indican que, en un ECOE virtual simulado de siete estaciones, ChatGPT obtuvo una puntuación media del 77,2%, superior a la puntuación histórica promedio de los candidatos humanos (73,7%), demostrando además una potencial capacidad para generar respuestas precisas, estructuradas y contextualmente relevantes ante escenarios clínicos complejos y dinámicos. El modelo fue superior a los candidatos humanos en diversas áreas de conocimiento, incluyendo el manejo del parto, la oncología ginecológica y la atención postoperatoria, al tiempo que mostró un alto desempeño en habilidades de comunicación, empatía y compromiso con el paciente. Asimismo, se distinguió por su capacidad para recuperar, integrar y sintetizar rápidamente información clínica compleja en respuestas coherentes, completando cada estación en un tiempo promedio de 2,54 minutos, muy por debajo de los 10 minutos asignados. De forma notable, algunos examinadores no pudieron distinguir entre las respuestas generadas por ChatGPT y las proporcionadas por los candidatos humanos.

En una línea complementaria, Li *et al.* [28] examinaron la precisión de los LLM para resolver estaciones de ECOE sobre asma en adultos, evaluando cinco iteraciones en tres escenarios clínicos estandarizados. Los resultados revelaron un desempeño moderado, con tasas de precisión situadas entre el 70% y el 85%, siendo ChatGPT-5 el modelo con la clasificación más alta entre los evaluados. Por su parte, Sethi *et al.* [29] evaluó el desempeño comparativo de diversos LLM (ChatGPT-3.5, Gemini Advanced y Claude Sonnet) frente a residentes de primer año de psiquiatría mediante evaluaciones estandarizadas que incluyeron exámenes teóricos y ECOE. Los resultados mostraron que el rendimiento de la inteligencia artificial en los ECOE varía considerablemente según el modelo utilizado. Mientras que Claude Sonnet fue superior a los residentes (20,0

Tabla 1. Características de los estudios incluidos sobre inteligencia artificial y estaciones del Examen Clínico Objetivo Estructurado.

Autor/ Año	País	Diseño	Mejor LLM	Porcentaje de precisión	Especialidad	Número de estaciones	Tiempo de resolución	Conclusiones
Sarah Li *et al.* 2023[24]	Singapur	Retrospectivo	Open AI (ChatGPT)	77,2%	Ginecología y Obstetricia	7	2,54 minutos	Los LLM superan a los humanos en responder las estaciones del ECOE.
Li *et al.* 2026[28]	Canadá	Experimental	Open AI (ChatGPT-5)	70-85%	Neumología	3	-	Los LLM muestran una alta precisión al responder estaciones de ECOE.
Sethi *et al.* 2026[29]	India	Transversal	Anthropic (Claude Sonnet)	93,4%	Psiquiatría	4	-	Los LLM superan a los residentes en las evaluaciones teóricas

GPT: generative pre-trained transformer; LLM: large language model.

frente a 16,6 puntos), ChatGPT-3.5 obtuvo puntuaciones inferiores (15,0 frente a 16,6 puntos).

Estos hallazgos sugieren que no existe un desempeño homogéneo de la IA en los ECOE y que la capacidad para resolver escenarios clínicos simulados está fuertemente influenciada por la arquitectura, el entrenamiento y las capacidades específicas de los modelos. Dichos modelos, principalmente ChatGPT-5 y Claude Sonnet, podrían constituir herramientas valiosas para su integración en futuras intervenciones educativas dirigidas tanto a pacientes como a profesionales de la salud.

Modelos de inteligencia artificial para generar estaciones del Examen Clínico Objetivo Estructurado

Si bien la IA ha demostrado una notable capacidad para generar y procesar grandes volúmenes de información, la evaluación de su potencial para diseñar exámenes tan complejos, rigurosos y estructurados como los ECOE representa una línea emergente de investigación. Los ECOE constituyen uno de los métodos de evaluación más utilizados en la formación médica para valorar competencias clínicas, comunicativas y de razonamiento diagnóstico. Sin embargo, su diseño e implementación requieren una considerable inversión de recursos humanos, tiempo y experiencia pedagógica. Asimismo, el desarrollo de estaciones y escenarios adaptados a distintas especialidades médicas plantea desafíos metodológicos y educativos significativos para los equipos docentes. En este contexto, la IA surge como una herramienta con el potencial de optimizar la creación de contenidos, apoyar el diseño de evaluaciones y ampliar las oportunidades de innovación en la educación médica basada en competencias.

Zouakia *et al.* evaluaron tres configuraciones de Generative Pre-trained Transformer (GPT) para la generación de estaciones de ECOE en salud digital [23]: un GPT estándar con instrucciones simples y pautas para ECOE; un GPT personalizado que añadió un libro de referencia en salud digital; y un GPT con agentes simulados, basado en instrucciones estructuradas que emulaban especialistas en ECOE e incorporaban el mismo libro. Se generaron 24 estaciones sobre ocho temas. El GPT con agentes simulados mostró el mejor cumplimiento del formato, calidad del contenido, precisión y preferencia general. El personalizado tuvo el menor cumplimiento del formato, mientras que el estándar obtuvo las puntuaciones más bajas en claridad y valor educativo. Estos hallazgos respaldan el uso de agentes expertos simulados e instrucciones estructuradas. El menor desempeño del GPT personalizado podría deberse a una sobrecarga cognitiva: añadir una fuente de conocimiento aumentó la complejidad de recuperación y dificultó priorizar las instrucciones clave.

Por otro lado, un estudio reciente realizado por Szychowiak *et al.* evaluaron la capacidad de los LLM, específicamente GPT-4o (OpenAI), para generar estaciones de ECOE listas para su implementación [30]. Siete expertos compararon cinco estaciones generadas por el modelo con cinco elaboradas por docentes. GPT-4o redujo la mediana del tiempo de elaboración de 54 a 24 minutos y generó resultados de laboratorio clínicamente relevantes. Sin embargo, presentó limitaciones para producir electrocardiogramas y radiografías de tórax adecuados. Mientras todas las estaciones docentes fueron calificadas como buenas, solo una generada por GPT-4o alcanzó ese nivel. Además, el modelo no siempre alineó las rúbricas con los datos clínicos y las instrucciones para los pacientes estandarizados. Por tanto, GPT-4o aún no genera consistentemente estaciones listas para su uso, por lo que la supervisión docente sigue siendo esencial.

La incorporación de sistemas de IA en la generación de estaciones de ECOE también podría contribuir a mejorar la estandarización y consistencia de las evaluaciones. En muchos contextos educativos, la elaboración de estaciones depende considerablemente de la experiencia individual de los docentes, lo que puede generar variabilidad en la complejidad, claridad y nivel de exigencia de los casos clínicos. El uso de modelos entrenados con criterios pedagógicos bien definidos podría facilitar la creación de escenarios más homogéneos y alineados con los objetivos de aprendizaje, reduciendo potenciales discrepancias entre estaciones y fortaleciendo la equidad del proceso evaluativo. En esta dirección, la evidencia disponible sugiere que los LLM, en particular los desarrollados por OpenAI, podrían tener la capacidad para reducir de manera sustancial el tiempo requerido en la creación de estaciones de ECOE, optimizando los recursos institucionales sin comprometer la calidad ni la validez de contenido de los escenarios generados. No obstante, estos modelos se encuentran en un proceso continuo de desarrollo y mejora, por lo que si se logra su implementación en un futuro debe acompañarse de una supervisión humana sistemática que permita garantizar la pertinencia pedagógica, la precisión de los contenidos y el cumplimiento de los estándares educativos establecidos.

Modelos de inteligencia artificial para evaluar las listas de verificación analíticas del Examen Clínico Objetivo Estructurado

La calificación de las listas de verificación en los ECOE representa uno de los mayores desafíos en la evaluación clínica estructurada, debido a su elevada dependencia de evaluadores humanos, la variabilidad interobservador y el considerable consumo de recursos institucionales. Ante este escenario, la aplicación de los LLM como herramientas de calificación automatizada

ha emergido como una de las líneas de investigación de mayor relevancia y proyección en la forma de evaluación médica contemporánea (Tabla 2).

El estudio de Sourial *et al.* realizado en Estados Unidos, evaluó la viabilidad del uso de la IA para calificar listas de verificación de ECOE en 39 estudiantes de Farmacia [31]. Mediante modelos de conversión de voz a texto y análisis de respuestas, la IA alcanzó una precisión superior al 96% y 94% en las estaciones A y B, respectivamente. Además, redujo el tiempo de evaluación de ocho horas a cinco minutos, aunque el tamaño de la muestra fue limitado. El estudio de Geathers *et al.* realizado en Estados Unidos, exploró el desempeño de cuatro grandes modelos de lenguaje (GPT-4o, Claude 3.5, Llama 3.1 y Gemini 1.5 Pro) para automatizar la evaluación de los ECOE mediante la MIRS (Master Interview Rating Scale, por sus siglas en inglés) [32]. Los modelos mostraron baja precisión exacta (0,27–0,52), pero una concordancia clínicamente útil con un margen de ±1 punto (0,67–0,91). No obstante, presentaron limitaciones para valorar la comunicación no verbal y aspectos sutiles de la relación médico-paciente. Por otro lado, el estudio transversal realizado por Tekin *et al.* [25] exploraron la concordancia entre evaluadores humanos y modelos de inteligencia artificial en la valoración de habilidades clínicas de estudiantes de medicina de primeros años durante los ECOE, utilizando el análisis de grabaciones de video. Las habilidades evaluadas incluyeron la administración de inyecciones intramusculares, la realización de nudos cuadrados, el soporte vital básico y el cateterismo urinario. El desempeño registrado en las grabaciones fue evaluado por cinco evaluadores: un evaluador humano en tiempo real, dos evaluadores humanos expertos que analizaron los videos posteriormente y dos sistemas basados en inteligencia artificial (ChatGPT-4o y Gemini Flash 1.5). La IA asignó puntuaciones generalmente mayores, excepto en cateterismo urinario, donde fueron similares. La concordancia fue superior en pasos visualmente observables y menor en tareas con componentes auditivos.

Un estudio reciente realizado por Luordo *et al.* [33], en España, exploró el uso del modelo de inteligencia artificial GPT-4 de OpenAI® junto con una instrucción diseñada por el equipo Digit-ECOE® para optimizar la precisión de los resultados en la calificación de informes clínicos elaborados por estudiantes durante el ECOE. Se analizaron las notas clínicas de 96 estudiantes de medicina, cuyos informes fueron evaluadas por dos expertos, un evaluador inexperto y la IA mediante una lista de verificación. GPT-4 fue más estricto y asignó puntuaciones promedio 3,51 puntos menores que los evaluadores humanos, pero redujo el tiempo de evaluación de 2–4 horas a 24 minutos. Aunque podría favorecer una calificación consistente y estandarizada, su limitada capacidad para reconocer variaciones válidas del lenguaje e inferencias implícitas puede generar discrepancias. A diferencia de la IA, los

Tabla 2. Características de los estudios incluidos sobre inteligencia artificial y evaluación de las estaciones del Examen Clínico Objetivo Estructurado.

Autor/Año	País	Diseño	LLM	Especialidad	Número de estudiantes	Número de estaciones	Tiempo de calificación (IA vs humano)	Conclusiones
Sourial *et al.* 2025[31]	EE. UU.	Cohorte	Inteligencia Artificial	Farmacia	39	2	5 minutos vs 8 horas	La IA superó la calificación del profesorado en las listas de verificación analíticas del ECOE.
Geathers *et al.* 2025[32]	EE. UU.	Experimental	GPT-4o Claude 3.5 Llama 3.1 Gemini 1.5 Pro	Medicina	-	-	-	LLM muestran un potencial considerable para la automatización eficaz y su aplicación en la evaluación de ECOE.
Tekin *et al.* 2025[25]	Turquía	Transversal	ChatGPT-4o Gemini Flash 1.5	Medicina	43	4	-	LLM muestran un potencial considerable para la automatización en la evaluación de ECOE
Luordo *et al.* 2025[33]	España	Cuasiexperimental	GPT-4	Medicina	96	1	24 minutos vs 2-4 horas	LLM es prometedora para la automatización en la evaluación de ECOE
Thomas *et al.* 2025[34]	EE. UU.	Comparativo	ChatGPT-4	Medicina	127	-	-	LLM califica de forma diferente en comparación con evaluadores humanos

ECOE: examen clínico objetivo estructurado; GPT: generative pre-trained transformer; LLM: large language model.

evaluadores humanos aplican su juicio clínico y suelen favorecer al estudiante ante respuestas aceptables, aunque no coincidan literalmente con los criterios establecidos.

El estudio de Thomas *et al.* comparó las calificaciones asignadas por ChatGPT con las otorgadas por un tutor humano al evaluar notas clínicas elaboradas por estudiantes de medicina [34]. Para ello, se analizaron 127 notas derivadas de un ECOE, las cuales contenían información relacionada con los antecedentes clínicos, el examen físico, el diagnóstico diferencial, el proceso de razonamiento clínico y el plan de tratamiento. Los resultados mostraron que las calificaciones asignadas por ChatGPT difirieron significativamente de las otorgadas por los tutores en los apartados de historia clínica, diagnóstico diferencial/proceso de razonamiento y plan de tratamiento. Según la evaluación de los tutores, 81 de las 127 notas (63,8%) fueron calificadas como sobresalientes y 46 (36,2%) como aprobadas. En contraste, ChatGPT otorgó una proporción significativamente mayor de calificaciones sobresalientes (118/127; 92,9%) y una menor proporción de aprobadas (9/127; 7,1%). Aunque ChatGPT-4 asignó calificaciones estadísticamente diferentes a las de los evaluadores humanos al analizar las notas clínicas tipo SOAP de los estudiantes, la magnitud práctica de estas diferencias fue relativamente pequeña. No obstante, la mayor discrepancia entre ambos métodos de evaluación se observó en la valoración del plan de tratamiento.

En caso de implementarse un modelo de IA en los ECOE de las universidades, uno de los principales beneficios sería la rapidez en la revisión y calificación de las evaluaciones. Un aspecto relevante es que muchos estudiantes experimentan ansiedad debido al tiempo de espera para conocer sus resultados [35]. Del mismo modo, los docentes enfrentan una considerable carga académica al tener que revisar un gran número de exámenes. Si se considera un escenario en el que un ECOE consta de ocho estaciones y es aplicado a 300 estudiantes por semestre, la carga de trabajo asociada a la corrección de las evaluaciones y a la entrega de resultados individuales representa un desafío significativo. En este contexto, las aplicaciones de IA podrían mejorar sustancialmente este proceso gracias a la rapidez con la que realizan las evaluaciones.

La automatización parcial de la calificación mediante IA podría representar una de las aplicaciones con mayor impacto en los ECOE. En la práctica, la evaluación de un gran número de estudiantes en un periodo corto de tiempo puede generar una considerable carga de trabajo para los docentes, aumentando el riesgo de inconsistencias en la aplicación de las rúbricas, errores involuntarios de registro o discrepancias entre el desempeño observado y la puntuación finalmente asignada. Asimismo, la interpretación subjetiva de algunos criterios de evaluación puede originar variabilidad entre evaluadores e incluso percepciones de inequidad por parte de los estudiantes. En este contexto, los sistemas de IA podrían contribuir a fortalecer la objetividad, estandarización y trazabilidad del proceso evaluativo, permitiendo una aplicación más uniforme de los criterios de calificación, una reducción de los tiempos de corrección y una entrega más oportuna de los resultados.

La evidencia disponible sugiere que la IA posee un importante potencial para complementar diversos componentes de los ECOE; sin embargo, su implementación no debería concebirse como un reemplazo de los evaluadores humanos, sino como una herramienta complementaria. Aspectos complejos como la empatía, la comunicación no verbal, el profesionalismo y las particularidades contextuales de la interacción clínica continúan requiriendo la interpretación y el juicio humano [21]. En consecuencia, un modelo híbrido que combine eficiencia, objetividad y capacidad analítica de la IA con la experiencia y criterio de los docentes podría representar la estrategia más adecuada para las evaluaciones clínicas del futuro.

Modelos de inteligencia artificial en la retroalimentación del Examen Clínico Objetivo Estructurado

Los ECOE son una evaluación de referencia en la formación médica; sin embargo, sus formatos tradicionales o virtuales suelen carecer de retroalimentación oportuna e individualizada, lo que puede limitar el desarrollo de competencias como la comunicación y la toma de decisiones. Uno de los principales beneficios de la IA es su capacidad para proporcionar retroalimentación. Los estudiantes suelen solicitar las respuestas correctas y comentarios sobre su desempeño, pues esto les permite identificar fortalezas y debilidades y tomar medidas para mejorar [36]. Aunque esta necesidad suele abordarse brevemente mediante reuniones grupales al finalizar los ECOE, muchos estudiantes requieren comentarios específicos basados en su propio examen. En un escenario con 300 estudiantes, ofrecer retroalimentación individualizada demandaría considerable tiempo y esfuerzo. Por ello, muchos docentes no la proporcionan, lo que genera críticas y frustración, especialmente entre quienes están insatisfechos con sus calificaciones. Los avances de la IA podrían contribuir a cubrir esta necesidad.

Dai *et al.* mediante un ensayo cruzado aleatorizado en 65 estudiantes de enfermería, evaluaron el impacto de la retroalimentación inmediata generada por inteligencia artificial (ChatGPT) e integrada en estaciones de ECOE basadas en realidad virtual sobre el desempeño comunicativo, la precisión en la toma de decisiones

clínicas y los resultados del aprendizaje [14]. Los resultados mostraron que la integración de retroalimentación inmediata generada por ChatGPT en escenarios de ECOE en realidad virtual mejoró significativamente las puntuaciones de comunicación, aumentó la precisión en la toma de decisiones clínicas y redujo el tiempo necesario para completar las evaluaciones. Además, se asoció con una mayor satisfacción con el aprendizaje, una mejor calidad de reflexión, una menor carga cognitiva y mejoras significativas en la autoeficacia académica y comunicativa. Estos hallazgos son consistentes con la teoría del aprendizaje experiencial, según la cual la reflexión iterativa y la resolución activa de problemas favorecen la consolidación de conocimientos y el desarrollo de habilidades. Otro estudio, realizado por Ali *et al.* evaluó el impacto de la retroalimentación generada por inteligencia artificial sobre el rendimiento académico y los niveles de ansiedad de los estudiantes durante un ECOE formativo [20]. Los estudiantes fueron asignados a grupos de intervención y control. El grupo de intervención recibió una capacitación integral sobre el uso de herramientas de IA (ChatGPT, Gemini y Perplexity) para generar materiales de estudio con retroalimentación personalizada más las instrucciones habituales del ECOE. El grupo de control solo recibió las instrucciones habituales del ECOE. Los resultados mostraron que no se encontró diferencia significativa entre los grupos de intervención y control con respecto a las calificaciones promedio del ECOE y en la puntuación total del cuestionario para la ansiedad.

Paralelamente, un reciente estudio de Borg *et al.* evaluó si la retroalimentación posterior a la consulta generada por inteligencia artificial e integrada en interacciones de realidad virtual con robótica social podía mejorar el desempeño clínico de 61 estudiantes de medicina [12]. Los resultados mostraron que los estudiantes que recibieron retroalimentación generada por IA obtuvieron puntuaciones totales significativamente más altas en los ECOE, así como tasas de aprobación significativamente superiores en comparación con aquellos que no recibieron esta intervención.

Estos hallazgos respaldan la posible integración de sistemas de retroalimentación basados en IA, previamente validados, como complemento de la enseñanza impartida por expertos durante las simulaciones virtuales orientadas a la formación clínica. Aunque algunos estudios no han identificado diferencias estadísticamente significativas en el desempeño de los estudiantes que reciben retroalimentación mediante IA, estos sistemas podrían constituir una herramienta complementaria para fortalecer los procesos de aprendizaje y evaluación en educación médica. En este contexto, futuros estudios con diseños metodológicos robustos y tamaños muestrales mayores deberían evaluar de manera prospectiva la efectividad de estas herramientas y determinar su impacto sobre desenlaces educativos.

Modelos de inteligencia artificial para mejorar habilidades en el Examen Clínico Objetivo Estructurado

La entrevista médica constituye una de las habilidades más importantes de la práctica clínica. Sin embargo, las oportunidades de formación práctica en las facultades de medicina suelen ser limitadas, principalmente debido a la escasez de hospitales y centros docentes, las dificultades de acceso a pacientes y la saturación de los campos clínicos. Esta situación ha mejorado considerablemente gracias a la estandarización de las prácticas y evaluaciones mediante los ECOE. No obstante, una de sus principales limitaciones es que los estudiantes generalmente solo tienen acceso a una o pocas sesiones de práctica por tema o especialidad médica durante cada rotación. Por ello, las aplicaciones basadas en inteligencia artificial podrían contribuir a cubrir estas deficiencias, mejorando tanto las oportunidades de práctica como las puntuaciones alcanzadas en los ECOE.

El ensayo controlado no aleatorizado realizado por Yamamoto *et al.* [26] proporcionó un programa de simulación basado en grandes modelos de lenguaje a 35 estudiantes japoneses, mientras 110 conformaron el grupo de control. Las puntuaciones del ECOE mostraron que el grupo que utilizó IA obtuvo resultados significativamente mayores en entrevistas médicas (media: 28,1; DE: 1,6 frente a 27,1; DE: 2,2; p = 0,01). Estos resultados sugieren que la formación mediante entrevistas con pacientes simulados por IA tiene un potencial educativo prometedor. Los participantes practicaron de forma realista con una IA capaz de simular respuestas emocionales, aproximándose a la interacción con pacientes reales. También se destacaron su accesibilidad y comodidad de uso. Sin embargo, se identificaron retrasos en la escritura y respuesta, errores del sistema y dificultades operativas, como el envío incorrecto de mensajes. A diferencia de la simulación tradicional con pacientes estandarizados, el uso de dispositivos portátiles permite aprender en un entorno psicológicamente seguro. Esto podría favorecer la comodidad dentro de la zona de confort, aunque posiblemente reducir su efectividad educativa. En cambio, cuando los estudiantes perciben la experiencia como desafiante, esta modalidad podría generar una verdadera zona de aprendizaje y potenciar la adquisición de conocimientos y habilidades. La integración de la simulación tradicional con pacientes estandarizados y las simulaciones mediante teléfonos inteligentes y computadoras podría potenciar el aprendizaje, mejorar la retroalimentación y contribuir a obtener mejores puntuaciones en los ECOE.

Otro ensayo aleatorizado realizado por Lee *et al*. evaluó y comparó la viabilidad y el impacto educativo de una simulación basada en un chatbot de inteligencia artificial frente a la dramatización tradicional entre pares (DTP) como estrategias de preparación para el ECOE [37]. Se aleatorizó a 19 estudiantes: nueve al grupo de chatbot y diez al de dramatización. El primer grupo practicó con una herramienta basada en GPT-4o y Claude 3.5, que proporcionaba respuestas contextualizadas según escenarios clínicos y retroalimentación automatizada. El segundo realizó prácticas en parejas bajo supervisión de un tutor. Luego, todos completaron dos estaciones de ECOE sobre mareo y dolor de hombro. No se observaron diferencias estadísticamente significativas entre ambos grupos. Los participantes de la DTP valoraron la autenticidad de las interacciones y la similitud con una situación real de examen. En contraste, quienes utilizaron el chatbot mostraron mayor satisfacción con la autonomía, las prácticas repetidas y la retroalimentación estructurada. La DTP parece eficaz para desarrollar habilidades procedimentales y comunicativas en contextos realistas, mientras que los chatbots podrían complementar el razonamiento clínico en un entorno reflexivo, flexible y adaptado al ritmo del estudiante, además de facilitar retroalimentación oportuna. Ambas modalidades presentan fortalezas complementarias, por lo que su integración podría constituir una estrategia prometedora para el desarrollo de competencias clínicas.

En conjunto, la integración del aprendizaje tradicional de los ECOE, grandes modelos de lenguaje y pacientes simulados basados en inteligencia artificial podrían potenciar el aprendizaje de los estudiantes y favorecer la obtención de puntuaciones más altas en los ECOE en comparación con los métodos de preparación tradicionales por sí sola.

BRECHAS POR RESOLVER Y DIRECCIONES FUTURAS

Las diversas aplicaciones de la IA en la evaluación mediante ECOE poseen un potencial transformador para la educación médica contemporánea, particularmente en lo que respecta a la resolución de limitaciones estructurales que afectan de manera sistemática la experiencia evaluativa de los estudiantes. En primer lugar, la latencia en la entrega de calificaciones, que en numerosos sistemas educativos puede extenderse varias semanas tras la realización del examen, constituye una fuente documentada de ansiedad académica y representa una barrera para el aprendizaje oportuno. En segundo lugar, la ausencia de retroalimentación individualizada limita la capacidad del estudiante para identificar sus errores específicos, reconocer sus fortalezas y orientar su desarrollo de competencias clínicas de forma autónoma y reflexiva. En tercer lugar, la variabilidad interobservador en la calificación introduce sesgos sistemáticos que comprometen la equidad del proceso evaluativo; estudios han documentado patrones de inequidad asociados a factores como el género del evaluador y del evaluado, reportándose, por ejemplo, que las examinadoras mujeres tendieron a otorgar puntuaciones más altas en términos generales, mientras que los examinadores hombres asignaron calificaciones significativamente más elevadas cuando la persona evaluada era mujer, así como sesgos vinculados a la pertenencia a minorías étnicas [38,39]. En cuarto lugar, las preferencias clínicas personales de los evaluadores influyen de manera sistemática en sus juicios, llevándolos a valorar conductas que ellos mismos priorizan en la práctica clínica, con independencia de los criterios establecidos en la lista de verificación, lo que configura un sesgo de contenido de difícil control bajo metodologías de evaluación convencionales [40,41]. Finalmente, la rigidez estructural de muchas estaciones de ECOE diseñadas para verificar el cumplimiento secuencial de listas de verificación no contempla adecuadamente la diversidad de razonamientos clínicos válidos ni la variabilidad en los procesos diagnósticos individuales, y el tiempo acotado de cada estación puede impedir una valoración integral de las competencias del estudiante [42]. Frente a estas brechas, las aplicaciones de la IA emergen como herramientas con capacidad para abordar de forma simultánea múltiples dimensiones del problema evaluativo; no obstante, su implementación efectiva requiere el desarrollo paralelo de programas de formación docente que garanticen el uso crítico, informado y éticamente responsable de estas tecnologías por parte de los evaluadores.

La integración de la inteligencia artificial en el enfoque de los ECOE abre una agenda de investigación de alta prioridad que trasciende la validación técnica de los modelos existentes. La investigación deberá abordar la equidad algorítmica de estos modelos, evaluando de forma sistemática la presencia de sesgos relacionados con el género, la etnia y el contexto sociocultural en las calificaciones generadas por IA, con el fin de asegurar que su adopción no reproduzca ni amplifique las inequidades documentadas en los sistemas de evaluación convencionales. Asimismo, el desarrollo de sistemas de retroalimentación automatizada en tiempo real capaces de analizar patrones de razonamiento clínico, identificar lagunas competenciales específicas y generar planes de mejora individualizados representa una dirección de alto impacto con implicaciones directas para la educación médica basada en la competencia. Finalmente, la incorporación de modelos multimodales con capacidad de procesamiento de audio, video y lenguaje no verbal abre la posibilidad de evaluar dimensiones relacionales y comunicativas de la competencia clínica que los sistemas actuales de listas de verificación son incapaces

de capturar con fidelidad. La implementación a escala de estas tecnologías exigirá marcos regulatorios y estándares de acreditación específicos que definan con precisión los roles, límites y responsabilidades de la IA en los procesos de certificación de competencias clínicas en las profesiones de la salud [11].

RECOMENDACIONES

Se recomienda la realización de futuros estudios, incluyendo ensayos clínicos y estudios observacionales, especialmente cohortes bien diseñadas, con el fin de explorar en mayor profundidad las diversas aplicaciones de la inteligencia artificial en el ámbito de los ECOE y la educación médica. Actualmente, numerosas universidades incorporan programas de simulación clínica cuya evaluación se basa en los ECOE. Por ello, es fundamental que estas instituciones reporten sistemáticamente sus resultados, experiencias y principales brechas identificadas, con el propósito de generar evidencia que permita orientar y fortalecer futuras investigaciones.

Las aplicaciones de la inteligencia artificial representan una de las áreas con mayor potencial de desarrollo para la ciencia contemporánea y, en consecuencia, su evolución debería progresar de manera paralela e integrada con los avances de la educación médica.

LIMITACIONES

Este estudio presenta varias limitaciones que deben considerarse al interpretar sus hallazgos. En primer lugar, la búsqueda bibliográfica se restringió a las bases de datos PubMed y Scopus, sin incluir otras fuentes relevantes como Web of Science, Embase o Education Resources Information Center (ERIC). Esta limitación pudo haber reducido la exhaustividad de la búsqueda y ocasionado que algunos estudios potencialmente elegibles no fueran identificados, lo que podría afectar la representatividad de la evidencia sintetizada. En segundo lugar, debido a que esta investigación corresponde a una revisión narrativa, no se aplicaron procedimientos metodológicos propios de las revisiones sistemáticas o de alcance, como un proceso estructurado de selección de estudios, estrategias de búsqueda exhaustivas o métodos estandarizados para la síntesis de la evidencia. Por tanto, los hallazgos deben interpretarse como una síntesis descriptiva de la literatura disponible y no como una estimación integral o definitiva del efecto de las aplicaciones de inteligencia artificial en los ECOE. En tercer lugar, la inclusión se limitó a publicaciones en inglés y español, lo que pudo haber excluido evidencia relevante publicada en otros idiomas y, potencialmente, introducir un sesgo de selección. Finalmente, no se realizaron evaluaciones formales del riesgo de sesgo, la calidad metodológica, la heterogeneidad ni la certeza de la evidencia de los estudios incluidos, debido a que el propósito de esta revisión fue proporcionar una síntesis narrativa de la literatura disponible. En consecuencia, la calidad y consistencia de los hallazgos individuales no pudieron ser ponderadas de manera sistemática, por lo que los resultados deben interpretarse con cautela y no permiten establecer conclusiones firmes sobre la efectividad, superioridad o aplicabilidad general de las herramientas de inteligencia artificial en los ECOE.

CONCLUSIONES

La inteligencia artificial puede contribuir a transformar la enseñanza y evaluación mediante ECOE al favorecer la estandarización, la retroalimentación oportuna y la reducción de la carga docente. Sin embargo, aún presenta limitaciones para valorar la comunicación no verbal, la empatía contextual y la flexibilidad del juicio clínico. Por ello, debe considerarse una herramienta complementaria, no un sustituto de la experiencia docente y clínica. Se requieren estudios rigurosos, con muestras adecuadas y desenlaces estandarizados, que evalúen su efectividad, seguridad, validez y aplicabilidad en distintos contextos educativos.

Material suplementario

Se puede acceder al material complementario en el siguiente enlace: https://revistas.cientifica.edu.pe/index.php/mecp/article/view/3854/1854

Contribución de autoría

NLCG: Conceptualización, Investigación, Administración del proyecto, Supervisión, Visualización, Validación, Redacción - borrador original; Redacción - revisión y edición. CCCH: Conceptualización, Curación de datos, Investigación, Visualización, Validación, Redacción - revisión y edición.

Financiamiento

Este estudio no ha recibido ningún tipo de financiamiento.

Declaración de conflictos de interés

Los autores declaran no tener conflictos de interés.

Declaración de disponibilidad de datos

No aplica.

Declaración de uso de inteligencia artificial generativa

Durante la elaboración de este trabajo, los autores utilizaron ChatGPT para mejorar la redacción, estilo

y gramática en algunas secciones del manuscrito. Posteriormente, los autores revisaron y editaron el contenido según fuera necesario y asumen toda la responsabilidad por el contenido del artículo publicado.

Agradecimientos

Ninguno.

ORCID

Nelson Luis Cahuapaza-Gutierrez: https://orcid.org/0000-0002-1228-1941

Cielo Cinthya Calderon-Hernandez: https://orcid.org/0009-0004-5242-3141

REFERENCIAS BIBLIOGRÁFICAS

1. Rayyan MR. The use of objective structured clinical examination in dental education- a narrative review. Front Oral Health. 2024;5:1336677. doi: 10.3389/froh.2024.1336677.
2. Barnes KN, Hardinger K, Hanson K, Gortney J. The Price of Preparedness: An Economic Evaluation of an OSCE in Pharmacy Training. Am J Pharm Educ. 2026;90(6):101990. doi: 10.1016/j.ajpe.2026.101990.
3. Weaver SB, Daftary M, Wingate L, Turner M. Evaluation of faculty inter-variability OSCE grade scoring on overall student performance in a laboratory course. Pharm Educ. 2022;22(1):48-53. doi: 10.46542/pe.2022.221.4853.
4. Sudan R, Clark P, Henry B. Cost and logistics for implementing the American College of Surgeons objective structured clinical examination. Am J Surg. 2015;209(1):140-4. doi: 10.1016/j.amjsurg.2014.10.001.
5. Dean J. A Golden Decade of Deep Learning: Computing Systems & Applications. Daedalus. 2022;151(2):58-74. doi: 10.1162/daed_a_01900.
6. Rejeleene R, Mehta NB. Artificial intelligence in medicine: How it works, how it fails. Cleve Clin J Med. 2026;93(2):113-20. doi: 10.3949/ccjm.93a.25089.
7. Occhipinti A, Verma S, Doan LMT, Angione C. Mechanism-aware and multimodal AI: beyond model-agnostic interpretation. Trends Cell Biol. 2024;34(2):85-9. doi: 10.1016/j.tcb.2023.11.002.
8. Kathiyan T, Kishore M, Priyanka D. An Intelligent AI-Based Framework for Automated Educational Video Generation. 2026 International Conference on Electronic Systems and Intelligent Computing (ICESIC), Chennai, India; 2026. pp. 373-378, doi: 10.1109/ICESIC67389.2026.11496457.
9. Woitsch R, Muck C, Utz W, Zeiner H. Enable Flexibilisation in FAIRWork's Democratic AI-based Decision Support System by Applying Conceptual Models Using ADOxx. Complex Systems Informatics and Modeling Quarterly. 2024;(38):27–53. doi: 10.7250/csimq.2024-38.02.
10. Flores-Núñez AM, Rivas-Sucari HC, Manrique-Rabelo CM, Rivas-Sucari SA, Flores-Núñez AM, Rivas-Sucari HC, et al. El impacto de la inteligencia artificial en la educación médica: aplicaciones, beneficios y desafíos. Gac Med Mex. 2026;162(2):194-8. doi: 10.24875/gmm.25000069.
11. Zafar I, Aboueisha H, Elhassan I, Shersad F, Caliskan SA, Syeda AF, *et al.* Ten tips for utilizing AI to generate high quality OSCE stations in medical education. Front Med (Lausanne). 2025;12:1744657. doi: 10.3389/fmed.2025.1744657.
12. Borg A, Schiött J, Ivegren W, Gentline C, Huss V, Hugelius AM, *et al.* AI-generated Feedback Following Social Robotic Virtual Patient Interactions and Medical Student Performance: Nonrandomized Quasi-Experimental Study. JMIR Med Educ. 2026;12:e90368. doi: 10.2196/90368.
13. Yokose M, Hirosawa T, Sakamoto T, Kawamura R, Suzuki Y, Harada Y, *et al.* The Validity of Generative Artificial Intelligence in Evaluating Medical Students in Objective Structured Clinical Examination: Experimental Study. JMIR Form Res. 2025;9:e79465. doi: 10.2196/79465.
14. Dai X, Gong Y, Ma LY. Effects of ChatGPT-generated immediate feedback integrated into VR-based OSCEs on nursing students' performance: a randomized crossover study. BMC Nurs. 2026;25:396. doi: 10.1186/s12912-026-04633-9.
15. Wang XJ, Song LJ, Jiao XP, Chen SQ. Integration of Artificial Intelligence in Nursing Clinical Education: Enhancing the Mini-CEX Model. J Multidiscip Healthc. 2025;18:7327-37. doi: 10.2147/JMDH.S550145.
16. Vázquez JT, Acosta HP, Loya JL. Integral evaluation of the medical graduate profile using the innovated ECOE-M. Educ Méd. 2026;27(3):101183. doi: 10.1016/j.edumed.2026.101183.
17. Chastain A, Schempp A. How Artificial Intelligence Can Affect Physician Assistant Student Self-Efficacy When Preparing for Objective Structured Clinical Examinations. J Physician Assist Educ. 2025;36(4):435-8. doi: 10.1097/JPA.0000000000000692.
18. Hicke Y, Geathers J, Vu K, Sewell J, Cardie C, Talwalkar J, *et al.* MedSimAI: Simulation and Formative Feedback Generation to Enhance Deliberate Practice in Medical Education. In: Proceedings of the LAK26: 16th International Learning Analytics and Knowledge Conference. 2026. doi: 10.1145/3785022.3785092.
19. Esmail S, Concannon B. Immersive Virtual Reality and AI (Generative Pretrained Transformer) to Enhance Student Preparedness for Objective Structured Clinical Examinations: Mixed Methods Study. JMIR Serious Games. 2025;13:e69428. doi:10.2196/69428.
20. Ali M, Rehman S, Cheema E. Impact of artificial intelligence on the academic performance and test anxiety of pharmacy students in objective structured clinical examination: a randomized controlled trial. Int J Clin Pharm. 2025;47(4):1034-41. doi:10.1007/s11096-025-01876-5.
21. Mishra GV, Luharia AA, Naqvi W, Sood A. Artificial Intelligence in OSCE: Innovations and Implications for Medical Education Assessment – A Systematic Review. In: 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI). 2024. doi: 10.1109/IDICAIEI61867.2024.10842789.
22. Jones S, Axman LM, Widemark E, Bafaloukos C, Sabado ET, Cranch-Kaniut C, *et al.* Enhancing the Objective Structured Clinical Examination Using Artificial Intelligence. J Dr Nurs Pract. 2025;JDNP-2025-0061.R1. doi: 10.1891/JDNP-2025-0061.
23. Zouakia Z, Logak E, Szymczak A, Jais JP, Burgun A, Tsopra R. AI-Driven Objective Structured Clinical Examination

Generation in Digital Health Education: Comparative Analysis of Three GPT-4o Configurations. JMIR Med Educ. 2026;12:e82116. doi: 10.2196/82116.

24. Li SW, Kemp MW, Logan SJS, Dimri PS, Singh N, Mattar CNZ, *et al.* ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol. 2023;229(2):172.e1-172.e12. doi: 10.1016/j.ajog.2023.04.020.
25. Tekin M, Yurdal MO, Toraman Ç, Korkmaz G, Uysal İ. Is AI the future of evaluation in medical education?? AI vs. human evaluation in objective structured clinical examination. BMC Med Educ. 2025;25(1):641. doi: 10.1186/s12909-025-07241-4.
26. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, *et al.* Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial. JMIR Med Educ. 2024;10:e58753. doi: 10.2196/58753.
27. Liu M, Okuhara T, Chang X, Shirabe R, Nishiie Y, Okada H, *et al.* Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. J Med Internet Res. 2024;26:e60807. doi: 10.2196/60807.
28. Li PY, Gershon A, Kouri A. Evaluating the Accuracy of Large Language Models in Answering Asthma Multiple Choice and Objective Structured Clinical Examination Questions. Chest. 2026;169(5):1286-97. doi: 10.1016/j.chest.2025.11.034.
29. Sethi MIS, Suhas S, Harbishettar V, Manjunatha N, Kumar CN, Bharadwaj R, *et al.* Benchmarking Large Language Models Against Psychiatry Residents Using Traditional Institutional Assessments. Indian J Psychol Med. 2026;02537176261435658. doi: 10.1177/02537176261435658.
30. Szychowiak P, Wong-So J, Messet H, Faure M, Breteau I, Jamard S, *et al.* Large language model-generated versus teacher-written objective structured clinical examination stations for medical students: a blinded comparative pilot study. J Educ Eval Health Prof. 2026;23:9. doi: 10.3352/jeehp.2026.23.9.
31. Sourial M, Hagler JC. A Pilot Study Using Artificial Intelligence to Enhance Efficiency, Accuracy, and Objectivity in Grading Pharmacy Objective Structured Clinical Examinations. Am J Pharm Educ. 2025;89(8):101455. doi: 10.1016/j.ajpe.2025.101455.
32. Geathers J, Hicke Y, Chan C, Rajashekar N, Young S, Sewell J, *et al.* Benchmarking Generative AI for Scoring Medical Student Interviews in Objective Structured Clinical Examinations (OSCEs). In: Cristea AI, Walker E, Lu Y, Santos OC, Isotani S, editors. Artificial Intelligence in Education. Cham: Springer Nature Switzerland; 2025. p. 231-45. doi: 10.1007/978-3-031-98420-4_17.
33. Luordo D, Arrese MT, Calvo CT, Shani KDS, Cruz LMR, Sánchez FJG, *et al.* Application of Artificial Intelligence as an Aid for the Correction of the Objective Structured Clinical Examination (OSCE). Appl Sci. 2025;15(3). doi: 10.3390/app15031153.
34. Thomas K, Szalacha L, Hanna K, Anibal J, Petrilli J. Evaluating the Effectiveness of ChatGPT Versus Human Proctors in Grading Medical Students' Post-OSCE Notes. Fam Med. 2025;57(10):727-31. doi: 10.22454/FamMed.2025.954255.
35. Alshareef N, Giga S, Fletcher I. Test anxiety, emotional regulation and academic performance among medical students: a qualitative study. Med Educ Online. 2025;30(1):2505177. doi: 10.1080/10872981.2025.2505177.
36. Daniels VJ, Ortiz S, Sandhu G, Lai H, Yoon MN, Bulut O, *et al.* Effect of Detailed OSCE Score Reporting on Learning and Anxiety in Medical School. J Med Educ Curric Dev. 2021;8:2382120521992323. doi: 10.1177/2382120521992323.
37. Lee HY, Kim J, Choi H, Bae H, Jeong A, Choi S, *et al.* Comparing AI chatbot simulation and peer role-play for OSCE preparation: a pilot randomized controlled trial. BMC Med Educ. 2025;25(1):1755. doi: 10.1186/s12909-025-08308-y.
38. Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, *et al.* Examiner effect on the objective structured clinical exam - a study at five medical schools. BMC Med Educ. 2017;17(1):71. doi: 10.1186/s12909-017-0908-1.
39. Claridge H, Stone K, Ussher M. The ethnicity attainment gap among medical and biomedical science students: a qualitative study. BMC Med Educ. 2018;18(1):325. doi: 10.1186/s12909-018-1426-5.
40. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B. The sights and insights of examiners in objective structured clinical examinations. J Educ Eval Health Prof. 2017;14:34. doi: 10.3352/jeehp.2017.14.34.
41. Hyde S, Fessey C, Boursicot K, MacKenzie R, McGrath D. OSCE rater cognition - an international multi-centre qualitative study. BMC Med Educ. 2022;22(1):6. doi: 10.1186/s12909-021-03077-w.
42. Myung SJ, Kim JW, Kim CW, Kim DH, Eo E, Kim JH, *et al.* Effect of limiting checklist on the validity of objective structured clinical examination: A comparative validity study. Med Teach. 2025;47(8):1360-6. doi: 10.1080/0142159X.2024.2430364.