Inteligencia artificial en la formación clínica práctica de estudiantes de medicina de pregrado: una revisión de alcance de aplicaciones, resultados y brechas.

RevEspEduMed **2026, 3, 710071; https://doi.org/10.6018.edumed.710071** revistas.um.es/edumed

7. Anexos.

Anexo 1. Estrategias de búsqueda bibliográfica.

Se presentan a continuación las estrategias completas utilizadas en PubMed/MEDLINE, Scopus y Web of Science (Core Collection), junto con el número de registros identificados en cada base de datos.

PubMed/MEDLINE: (("Artificial Intelligence"[Mesh] OR "Machine Learning"[Mesh] OR "Deep Learning"[Mesh] OR "artificial intelligence"[tiab] OR "machine learning"[tiab] OR "deep learning"[tiab] OR "large language model*"[tiab] OR LLM*[tiab] OR "generative AI"[tiab] OR ChatGPT[tiab] OR chatbot*[tiab] OR "intelligent tutor*"[tiab] OR "clinical decision support"[tiab] OR "decision support system*"[tiab]) AND ("Students, Medical"[Mesh] OR "Education, Medical, Undergraduate"[Mesh] OR "medical student*"[tiab] OR "undergraduate medical education"[tiab] OR "medical education"[tiab] OR "medical school*"[tiab] OR clerkship*[tiab] OR "clinical clerkship*"[tiab]) AND ("Clinical Competence"[Mesh] OR "clinical reasoning"[tiab] OR "clinical decision*"[tiab] OR "diagnostic reasoning"[tiab] OR "diagnostic accuracy"[tiab] OR OSCE[tiab] OR "objective structured clinical examination"[tiab] OR "mini-CEX"[tiab] OR EPA*[tiab] OR "entrustable professional activit*"[tiab] OR simulation[tiab] OR "clinical skill*"[tiab] OR "clinical competenc*"[tiab] OR "clinical performance"[tiab]) AND (education*[tiab] OR learn*[tiab] OR teaching[tiab] OR training[tiab] OR curriculum[tiab] OR instruction*[tiab])) NOT ("model performance"[tiab] OR "algorithm performance"[tiab] OR "diagnostic model"[tiab] OR "predictive model"[tiab] OR "neural network"[tiab] OR "classification performance"[tiab] OR ROC[tiab] OR AUC[tiab])

Registros identificados: 793

Scopus: TITLE-ABS-KEY(("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model*" OR "generative AI" OR ChatGPT OR chatbot* OR "intelligent tutor*" OR "clinical decision support") AND ("medical student*" OR "undergraduate medical education" OR "medical school*" OR clerkship* OR "clinical clerkship*") AND ("clinical reasoning" OR "diagnostic reasoning" OR "clinical decision*" OR OSCE OR "objective structured clinical examination" OR "mini-CEX" OR EPA* OR simulation OR "clinical competenc*" OR "clinical performance") AND (education* OR learning OR teaching OR training OR curriculum OR instruction*) AND (assess* OR evaluat* OR outcome* OR "performance" OR "skills")) AND NOT TITLE-ABS-KEY ("model performance" OR "algorithm performance" OR "classification performance" OR "predictive model" OR "diagnostic model" OR "feature extraction" OR "convolutional neural network" OR "support vector machine")

Registros identificados: 529

Web of Science (Core Collection): TS=(("artificial intelligence" OR "machine learning" OR "deep learning" OR "large language model*" OR LLM* OR "generative AI" OR ChatGPT OR chatbot* OR "intelligent tutor*" OR "clinical decision support" OR "decision support system*") AND ("medical student*" OR "undergraduate medical education" OR "medical education" OR "medical school*" OR clerkship* OR "clinical clerkship*")

AND ("clinical reasoning" OR "clinical decision*" OR "diagnostic reasoning" OR "diagnostic accuracy" OR OSCE OR "objective structured clinical examination" OR "mini-CEX" OR EPA* OR "entrustable professional activit*" OR simulation OR "clinical skill*" OR "clinical competenc*" OR "clinical performance") AND (education* OR learning OR teaching OR training OR curriculum OR instruction*) NOT ("model performance" OR "algorithm performance" OR "classification performance" OR "diagnostic model" OR "predictive model"))

Registros identificados: 790

Tabla 1. Criterios de inclusión y exclusión utilizados.

Criterios de Inclusión	Criterios de Exclusión
Población Objeto de Estudio	
Estudios que involucren a estudiantes de pregrado de medicina (MD/MBBS o equivalente).	Estudios enfocados exclusivamente en médicos residentes, especialistas u otros profesionales de la salud.
Estudios con población mixta, se incluyen si presentan resultados diferenciables para el nivel de pregrado.	
Intervención	
Intervenciones educativas que incorporen Inteligencia Artificial definida explícitamente (aprendizaje automático/ML, aprendizaje profundo/DL, modelos de lenguaje grande/LLM, IA generativa, procesamiento de lenguaje natural/NLP, visión por computadora, algoritmos entrenados), utilizada para la enseñanza, entrenamiento, evaluación o retroalimentación de competencias clínicas.	Estudios sobre intervenciones educativas médicas que no integren Inteligencia Artificial.
Contexto	
Estudios relativos a contextos de formación práctica: práctica clínica real bajo supervisión y/o simulación (Examen Clínico Objetivo Estructurado/OSCE, Laboratorio de destrezas/skills lab, Pacientes virtuales con IA, casos simulados interactivos).	
Diseño y Resultados	
Estudios de naturaleza empírica (cuantitativos, cualitativos o mixtos) que reporten resultados vinculados al aprendizaje, al desempeño o a aspectos sustantivos de implementación educativa.	Estudios cuya evaluación se centre exclusivamente en satisfacción, percepción o actitudes hacia la Inteligencia Artificial, sin reportar resultados vinculados al aprendizaje, al desempeño o a la implementación educativa.

	Estudios en los cuales la Inteligencia Artificial sea empleada para fines distintos a los educativos.
Tipo de Publicación	
	Editoriales, comentarios, cartas al editor, ensayos de opinión, protocolos de estudio sin resultados publicados y revisiones secundarias.
Idioma y Periodo Temporal	
Idiomas aceptados: inglés, español y portugués.	Cualquier otro idioma.
Publicaciones a partir del año 2021 (con justificación fundamentada).	

Tabla 2. Resumen descriptivo de los hallazgos de los estudios incluidos

Autores y Año	País	Diseño del Estudio	Participantes (*n*)	Tecnología de IA	Competencia clínica principal	Resultados Principales
Yamamoto et al. (2024)	Japón	Ensayo controlado no aleatorizado	35 estudiantes de medicina	GPT-4 (paciente simulado)	Entrevista clínica / historia clínica y comunicación	El estudio reportó puntuaciones más altas en las evaluaciones de entrevistas médicas del OSCE en la condición con paciente simulado basado en IA.
Ali et al. (2022)	Canadá	Ensayo clínico aleatorizado (ECA)	70 estudiantes	Virtual Operative Assistant (VOA)	Habilidades técnicas/procedimentales	El estudio reportó mayores puntuaciones de habilidad quirúrgica en la situación con tutor de IA que en los basados en instrucción humana.
Brügge et al. (2024)	Alemania	ECA doble ciego	21 estudiantes	ChatGPT 3.5	Razonamiento clínico / CDM	El estudio reportó mayores puntuaciones de toma de decisiones clínicas en el grupo que recibió feedback estructurado por IA respecto del que no recibió feedback por IA.
Holderried et al. (2024)	Alemania	Estudio prospectivo	106 interacciones	GPT-4	Evaluación/ feedback automatizado (assessment-centric)	El estudio encontró una concordancia "casi perfecta" (κ = 0.832) entre el feedback automatizado generado por IA y las evaluaciones humanas.
Giglio et	Canadá	ECA de 3	88	ICEMS	Habilidades	El estudio reportó mejor

al. (2025)		brazos (ciego simple)	estudiantes	(Tutor de IA)	técnicas/proce dimentales	desempeño en riesgos específicos de seguridad quirúrgica en la condición con instrucción experta aumentada por IA.
Gomez et al. (2026)	Suiza	Estudio de usuario y evaluación	5 expertos y 12 estudiantes	YOLOX-S + MSTCN++	Habilidades técnicas/proce dimentales	El estudio evaluó un sistema de IA explicable (XAI) para la evaluación automatizada de movimientos durante tareas de sutura.
Ba et al. (2024)	China	Diseño experimen tal controlado	77 internos	ChatGPT 4.0	Competencia clínica integrada (Mini-CEX)	El estudio reportó mayores puntuaciones de competencia clínica y comunicación médico-paciente en la condición con apoyo de IA.
Li & Zhao (2026)	China	ECA prospectiv o (ciego simple)	110 estudiantes	PACS+AI (scaffoldin g cognitivo)	Imagenología e interpretación diagnóstica	El estudio reportó mayor retención de conocimientos y precisión diagnóstica a los 3 meses en la condición con PACS asistido por IA.
Lippitsch et al. (2024)	Alemani a	ECA controlado	168 entrevistas	AIML (ViPATalk)	Entrevista/ historia clínica	El estudio reportó resultados equivalentes o no inferiores para la práctica con IA en comparación con el juego de roles tradicional entre pares.
Liu et al. (2025)	China	Estudio prospectiv o multicase	31 estudiantes	DeepSeek-V2.5 (AMTES)	Evaluación/ feedback automatizado (assessment-centric)	El estudio encontró alta estabilidad y consistencia entre la evaluación automatizada por IA y la evaluación humana en tareas de anamnesis (ICC > 0.923).
Luo et al. (2025)	China	ECA prospectiv o	84 estudiantes	Baichuan-13B (LLMDP)	Entrevista/ historia clínica	El estudio reportó un incremento de 10.5 puntos en habilidades de anamnesis oftalmológica y mayor empatía percibida en la condición con paciente digital.
Co et al. (2022)	Hong Kong	Estudio de casos y controles	132 estudiantes	Dialogflo w ("Bennie and	Entrevista clínica / historia clínica y	El estudio reportó desempeño comparable entre el grupo con chatbot y el grupo convencional en

				Chats")	comunicación	habilidades de anamnesis, junto con resultados de usabilidad y eficacia percibida del aprendizaje.
Öncü et al. (2025)	Turquía	Diseño de triangulación simultánea	21 estudiantes	ChatGPT-4o	Competencia clínica integrada (OSCE/Mini-CEX/múltiples subdominios)	El estudio exploró el uso de ChatGPT-4o para identificar insuficiencias en la gestión de casos y en la resolución de problemas bajo presión.
Kıyak et al. (2025)	Turquía	Estudio prospectivo paralelo	71 estudiantes	GPT-4o-mini	Razonamiento clínico / diagnóstico	El estudio reportó que los estudiantes de primer año con apoyo de IA obtuvieron mejores resultados que el grupo comparador en razonamiento diagnóstico para los casos entrenados.
Sun & Liu (2025)	China	ECA controlado	124 estudiantes	MetaGP-SurgEd (Video-LLM)	Habilidades técnicas/procedimentales (quirúrgicas)	El estudio reportó mejoras en eficiencia procedimental y en puntuaciones de habilidades quirúrgicas en la condición con tutoría basada en IA.
Sun & Liu (2025)	China	Cohorte retrospectiva multicéntrica	1632 estudiantes	MetaGP-Edu	Razonamiento clínico y toma de decisiones diagnósticas	El estudio encontró que el acceso a la herramienta de IA se asoció con un aumento promedio de 8.2 puntos en las notas de Medicina Interna.
Turner et al. (2025)	EE. UU.	Estudio piloto	176 estudiantes	GPT-4 (Plataforma 2-Sigma)	Razonamiento clínico y toma de decisiones diagnósticas	El estudio reportó una precisión diagnóstica del 84% en más de 1.600 sesiones simuladas de casos cardiopulmonares.
Sheth et al. (2025)	Canadá	Análisis cuali-cuantitativo	94 participantes	OSCEai (IA generativa)	Entrevista clínica / historia clínica y comunicación	El estudio reportó un aumento en la comodidad del estudiante con la anamnesis, con un tamaño de efecto grande (d = 1.13).
Wang et al. (2025)	China	ECA de centro único	56 estudiantes	GPT-4 (GPT personalizado)	Entrevista clínica / historia clínica y comunicación	El estudio reportó puntuaciones más altas en la condición de entrenamiento basada en GPT que en el método tradicional.

Xie et al. (2025)	China	Estudio de rotación clínica	97 estudiantes	Multi-condition Diffusion Model	Imagenología e interpretación diagnóstica	El estudio reportó mejora en la precisión diagnóstica de queratitis mediante el uso de imágenes sintéticas generadas por IA.
Yajun Chen (2025)	China	ECA prospectivo	40 estudiantes	Coze (Plataforma personalizada)	Razonamiento clínico y toma de decisiones diagnósticas	El estudio reportó mejora en la adquisición de conocimientos (d = 0.72) y en la participación activa en la condición con plataforma personalizada basada en IA.
Lau et al. (2025)	Singapur	Ensayo controlado	No reportado	Kosmos (Ecógrafo con IA)	Imagenología e interpretación diagnóstica	El estudio describió el uso de un ecógrafo con IA que proporcionó guía en tiempo real y sugerencias para mejorar la calidad de las imágenes ecográficas.
Hui et al. (2025)	China	Estudio comparativo de PBL	42 internos	ChatGPT 4.0	Competencia clínica integrada	El estudio reportó puntuaciones más altas en exámenes teóricos y en competencia clínica general en la condición con PBL asistido por ChatGPT.
Zheng et al. (2024)	China	Diseño experimental controlado	66 estudiantes	ChatGPT + Midjourney	Competencia clínica integrada (OSCE/Mini-CEX/múltiples subdominios)	El estudio reportó mejora en pensamiento crítico clínico y mayor satisfacción con el contenido en la condición con simulación apoyada por IA.

Tabla 3. Estudios con comparación directa (IA vs comparador)

Estudio (Autores y Año)	Intervención de IA	Método Tradicional / Control	Competencia Evaluada	Hallazgo comparativo principal reportado
Yamamoto et al. (2024)	Paciente simulado por GPT-4.	Programa educativo tradicional.	Habilidades de entrevista médica (OSCE).	El estudio reportó puntuaciones más altas en habilidades de entrevista médica en el grupo con paciente simulado por GPT-4 que en el programa educativo tradicional (28.1 vs 27.1; p = .01).
Ali et al. (2022)	Tutor de IA (Virtual Operative Assistant).	Instrucción de experto remoto / Sin feedback.	Pericia en neurocirugía (resección de tumores).	El estudio reportó mayores puntuaciones de pericia en neurocirugía en el grupo con tutor de IA que en los grupos con instrucción de experto remoto o sin feedback (p < .001).
Sun & Liu (2025) (Quirúrgico)	Tutor de IA (MetaGP-SurgEd).	Instrucción por cirujanos expertos / Autoaprendizaje.	Habilidades laparoscópicas fundamentales.	El estudio reportó resultados comparables entre el tutor de IA y la instrucción por expertos humanos, y superiores al autoaprendizaje en habilidades laparoscópicas fundamentales (p < 0.001).
Li & Zhao (2026)	Andamiaje cognitivo en PACS+AI.	Flujo de trabajo PACS estándar sin guía.	Conocimiento y decisión clínica en radiología.	El estudio reportó mayor retención de conocimientos a los 3 meses en el grupo con andamiaje cognitivo en PACS+IA que en el flujo de trabajo PACS estándar sin guía (79.3% vs 19.7%; p < 0.001).
Luo et al. (2025)	Paciente digital (LLMDP) basado en Baichuan-13B.	Entrenamiento tradicional con pacientes reales.	Anamnesis en oftalmología.	El estudio reportó un incremento promedio de 10.5 puntos en las evaluaciones de historia clínica en la condición con paciente digital, en comparación con el

				entrenamiento tradicional con pacientes reales ($p < 0.001$).
Wang et al. (2025)	Entrenamiento con un GPT personalizado.	Método tradicional de juego de roles (role-playing).	Recolección de historia y razonamiento clínico.	El estudio reportó puntuaciones de examen más altas en el grupo con GPT personalizado que en el método tradicional de juego de roles (86.79 vs 73.64; $p < 0.001$).
Ba et al. (2024)	Instrucción asistida por ChatGPT-4.0.	Enseñanza tradicional al lado de la cama (bedside teaching).	Competencia clínica en pediatría (Mini-CEX).	El estudio reportó mayores puntuaciones en comunicación y juicio clínico en la condición con instrucción asistida por ChatGPT-4.0, sin diferencias en conocimiento teórico frente a la enseñanza tradicional al lado de la cama ($p < 0.05$).
Hui et al. (2025)	PBL asistido por ChatGPT.	Enseñanza tradicional.	Conocimiento teórico y habilidades en urología.	El estudio reportó puntuaciones más altas en examen teórico y Mini-CEX en el grupo con PBL asistido por ChatGPT que en la enseñanza tradicional ($p < 0.01$ para ambos outcomes).
Zheng et al. (2024)	Simulación basada en escenarios con IA.	Modo de enseñanza tradicional.	Conocimiento y pensamiento crítico en cardiología.	El estudio reportó puntuaciones más altas en teoría y habilidades clínicas en la condición con simulación basada en escenarios con IA que en el modo de enseñanza tradicional ($p < 0.001$ y $p < 0.05$, respectivamente).
Kıyak et al. (2025)	Preguntas y feedback por GPT-4o-mini.	Currículo estándar (estudiantes de nivel superior como control).	Razonamiento diagnóstico quirúrgico.	El estudio reportó que los estudiantes de primer año con preguntas y feedback por GPT-4o-mini obtuvieron mejores resultados en diagnósticos entrenados que el grupo comparador sin IA ($p < 0.001$).

Sun & Liu (2025) (Medicina)	Acceso a la herramienta MetaGP-Edu.	Cohorte histórica con enseñanza tradicional.	Razonamiento clínico en Medicina Interna.	El estudio encontró que la disponibilidad de MetaGP-Edu se asoció con un incremento promedio de 8.2 puntos en la nota final del curso, en comparación con una cohorte histórica con enseñanza tradicional ($p < 0.001$).
Lau et al. (2025)	Ecógrafo con guía de IA (Kosmos).	Instrucción tradicional presencial.	Adquisición de imágenes en ecografía cardíaca.	El estudio reportó no inferioridad del entrenamiento con ecógrafo guiado por IA frente a la instrucción tradicional presencial para adquisición de imágenes en ecografía cardíaca ($p < 0.001$).
Lippitsch et al. (2024)	Sistema de avatares ViPATalk.	Juego de roles tradicional entre pares (peer role play).	Completitud de la anamnesis médica.	El estudio reportó resultados mayormente equivalentes o no inferiores para el sistema de avatares ViPATalk en comparación con el juego de roles tradicional entre pares en la completitud de la anamnesis médica.
Co et al. (2022)	Chatbot móvil (Bennie and Chats).	Enseñanza tradicional presencial.	Habilidades de anamnesis quirúrgica.	El estudio reportó desempeño comparable entre el chatbot móvil y la enseñanza tradicional presencial en habilidades de anamnesis quirúrgica ($p > 0.05$).
Xie et al. (2025)	Imágenes generadas por IA generativa.	Casos reales e imágenes médicas reales.	Diagnóstico de queratitis infecciosa.	El estudio reportó resultados comparables entre el entrenamiento con imágenes generadas por IA y el uso de imágenes médicas reales para el diagnóstico de queratitis infecciosa; adicionalmente, el grupo con IA superó a modelos previos de clasificación.
Brügge et al. (2024)	Feedback estructurado por	Simulación con IA pero sin feedback	Toma de decisiones clínicas (CDM).	El estudio reportó mayores puntuaciones de

	ChatGPT.	adicional (control).		toma de decisiones clínicas en el grupo con feedback estructurado por ChatGPT que en la condición control sin feedback adicional (p = .049), con un tamaño de efecto elevado.
Gomez et al. (2026)	Feedback de IA Explicable (XAI).	Coaching tradicional basado en video.	Habilidades de sutura quirúrgica.	El estudio no encontró diferencias estadísticamente significativas entre el feedback de IA explicable (XAI) y el coaching tradicional basado en video en habilidades de sutura quirúrgica (p = 0.254), aunque describió tendencias favorables en el grupo XAI.

Nota formal: comparador tradicional/humano, histórico, o condición alternativa con IA