Revista Española
de Educación Médica

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

Artificial intelligence in the practical clinical training of undergraduate medical students: a scoping review of applications, outcomes, and gaps.

Juan Bonilla[1], Tomás Cortés[2], Diego Polanco[3], Carlos Martínez[4], Álvaro Herrera[5]

1, Escuela de Medicina, Facultad de Medicina, Universidad de Chile. Santiago, Chile; juan.mejia@ug.uchile.cl, ORCID-ID 0009-0007-4086-083X; 2, tomas.cortes.f@ug.uchile.cl, ORCID-ID 0009-0008-4768-2746; 3, diego.polanco@ug.uchile.cl, ORCID-ID 0009-0008-9502-465X; 4, carlosmartinez@ug.uchile.cl ORCID-ID 0009-0001-1136-1026; 5, Departamento de Educación en Ciencias de la Salud, Facultad de Medicina, Universidad de Chile. Santiago, Chile; levarito@uchile.cl, ORCID-ID 0009-0007-4861-2144
* Correspondencia: juan.mejia@ug.uchile.cl

Recibido: 8/4/26; Aceptado:4/5/26; Publicado: 6/5/26

Resumen.

Objetivo: Mapear y sintetizar la evidencia disponible sobre el uso de inteligencia artificial (IA), incluidos los modelos de lenguaje de gran escala (LLM) y las herramientas generativas, en la formación clínica práctica de estudiantes de medicina de pregrado. **Metodología:** Se realizó una scoping review siguiendo el marco metodológico de Arksey y O'Malley, y se reportó de acuerdo con PRISMA-ScR. La búsqueda bibliográfica se llevó a cabo el 28 de enero de 2026 en PubMed/MEDLINE, Scopus y Web of Science Core Collection. Se incluyeron estudios empíricos publicados desde 2021 en adelante, en inglés, español o portugués, que evaluaran intervenciones educativas basadas en IA en estudiantes de medicina de pregrado, en contextos de formación práctica clínica supervisada en escenarios reales y/o simulados. **Resultados:** Se identificaron 2112 registros, de los cuales 789 fueron eliminados por duplicación. Tras el cribado y la evaluación de texto completo, se incluyeron 24 estudios. La evidencia se concentró en escenarios simulados o estructurados y en dominios como entrevista clínica/comunicación, razonamiento clínico y habilidades técnicas o procedimentales. Los LLM y las herramientas generativas fueron las tecnologías más frecuentemente estudiadas. Un subconjunto de estudios comparativos reportó resultados comparables o, en determinados dominios y contextos específicos, favorables para intervenciones basadas en IA; sin embargo, la heterogeneidad metodológica de los comparadores, outcomes y diseños impide extraer conclusiones agregadas sobre efectividad, y la evidencia se concentró principalmente en desenlaces educativos inmediatos. **Conclusiones:** La IA muestra potencial como herramienta complementaria para ampliar la práctica deliberada, estandarizar la retroalimentación y apoyar experiencias de aprendizaje más accesibles y, en algunos casos, más personalizadas. No obstante, persisten limitaciones relacionadas con la heterogeneidad metodológica, la escasa evaluación en contextos clínicos reales y la falta de seguimiento longitudinal, por lo que se requieren estudios más robustos y marcos éticos y pedagógicos claros que orienten su integración responsable en la educación médica de pregrado.

Keywords: Inteligencia artificial, Modelos de Lenguaje a Gran Escala, Educación médica, Competencia clínica, Estudiantes de medicina, Simulación clínica, Revisión de alcance.

Abstract.

Objective: To map and synthesize the available evidence on the use of artificial intelligence (AI), including large language models (LLMs) and generative tools, in the practical clinical training of undergraduate medical students. **Methods:** A scoping review was conducted following the Arksey and O'Malley methodological framework and reported in accordance with PRISMA-ScR. The literature search was carried out on January 28, 2026, in PubMed/MEDLINE, Scopus, and Web of Science Core Collection. Empirical studies published from 2021 onward in English, Spanish, or Portuguese were included if they evaluated AI-based educational interventions for undergraduate medical students in supervised practical clinical training in real and/or simulated settings. **Results:** A total of 2,112 records were identified, of which 789 were removed as duplicates. After screening and full-text assessment, 24 studies were included. The evidence was concentrated in simulated or structured settings and in domains such as clinical interviewing/communication, clinical reasoning, and technical or procedural skills. LLMs and generative tools were the most frequently studied technologies. A subset of comparative studies reported outcomes that were comparable or, in specific domains and contexts, favorable to AI-based interventions; however, the methodological heterogeneity of comparators, outcomes, and study designs precluded drawing aggregated conclusions about effectiveness, and the evidence was focused mainly on immediate educational outcomes. **Conclusions:** AI shows potential as a complementary tool to expand deliberate practice, standardize feedback, and support more accessible and, in some cases, more personalized learning experiences. Nevertheless, important limitations remain, including methodological heterogeneity, limited evaluation in real clinical settings, and the lack of longitudinal follow-up. More robust studies and clear ethical and pedagogical frameworks are needed to guide its responsible integration into undergraduate medical education.

Keywords: Artificial intelligence, Large-scale language models, Medical education, Clinical competence, Medical students, Clinical simulation, Scope review.

---

1. Introducción

La formación clínica práctica en medicina busca que el estudiante desarrolle competencias observables en contextos reales o simulados—por ejemplo, entrevista clínica, examen físico, comunicación, razonamiento clínico aplicado y desempeño en escenarios tipo OSCE o estaciones de habilidades–. Sin embargo, la enseñanza y evaluación de estas competencias suele ser exigente en término de los recursos necesarios (tiempo docente, evaluadores, pacientes estandarizados, infraestructura, entre otros) y, por tanto, difícil de escalar de forma consistente a cohortes crecientes (1). En paralelo, los currículos basados en competencias han aumentado la demanda por práctica deliberada y retroalimentación frecuente y estandarizada a lo largo del pregrado.

En los últimos años, la inteligencia artificial (IA), incluyendo aprendizaje automático, sistemas adaptativos y, más recientemente, modelos generativos y modelos de lenguaje (LLM)—ha comenzado a incorporarse en educación médica con fines que van desde apoyo al aprendizaje y evaluación hasta simulación y entrenamiento de habilidades; revisiones de alcance previas han mapeado aplicaciones de IA de manera amplia donde destaca el entrenamiento de habilidades técnicas quirúrgicas, la evaluación automatizada/objetiva, las herramientas para el apoyo del razonamiento diagnóstico, entre otras (p. ej., admisiones, docencia, entrenamiento en laboratorios), destacando tanto oportunidades como la necesidad de marcos éticos y de gobernanza educativa (2-3). No obstante, estas síntesis suelen abarcar múltiples dominios y niveles formativos, por lo que la evidencia específica sobre IA aplicada a la formación exclusivamente práctica de estudiantes de medicina permanece dispersa.

Particularmente, la irrupción de herramientas basadas en LLM ha impulsado nuevas propuestas para entrenamiento práctico escalable, como pacientes simulados conversacionales y sistemas que buscan automatizar componentes del feedback y la evaluación (4-5). Por ejemplo, se han descrito sistemas de pacientes simulados impulsados por agentes basados en LLM orientados a replicar interacciones clínico-comunicacionales, aunque su efectividad, confiabilidad y condiciones de implementación siguen siendo desafíos centrales (4, 6). En la misma línea, han surgido aplicaciones de IA para entrenamiento/evaluación en formatos tipo OSCE (p. ej., chatbots para entrevista clínica; evaluación automatizada del desempeño) (3). A pesar de este crecimiento, aún no está claro qué tipos de herramientas se están usando, en qué escenarios prácticos, con qué diseños de evaluación, qué outcomes competenciales se miden y qué brechas metodológicas y éticas persisten, especialmente considerando la variabilidad en definiciones de "IA" y la limitada evidencia robusta en outcomes educativos/clínicos señalada por síntesis previas. En este contexto, comprender de manera sistemática cómo se está utilizando la IA en la formación clínica práctica es especialmente relevante, dado que estas competencias impactan directamente en la calidad y seguridad de la atención médica futura. Además, la presión por escalar la enseñanza práctica en escenarios de recursos limitados hace que las soluciones basadas en IA resulten atractivas, pero su adopción requiere una caracterización clara de su alcance, desempeño reportado y posibles implicancias éticas y pedagógicas.

Para efectos de esta revisión, la formación clínica práctica se entiende de manera amplia, como un continuo de escenarios orientados al desarrollo y evaluación de competencias clínicas aplicadas, que abarca desde la práctica clínica real supervisada hasta contextos estructurados de simulación, incluyendo OSCE, Mini-CEX, laboratorios de destrezas, pacientes virtuales o digitales y casos simulados interactivos. Sin embargo, estos escenarios no se asumen como equivalentes entre sí, ya que difieren en su grado de autenticidad pedagógica, complejidad contextual y potencial de transferencia al desempeño clínico con pacientes reales. Esta distinción es relevante para interpretar la evidencia disponible, particularmente en un campo donde las aplicaciones de IA tienden a concentrarse en entornos más estructurados y simulables.

Dado que se trata de un campo emergente, heterogéneo en tecnologías, contextos educativos y métricas de evaluación, el objetivo de este estudio es mapear, explorar y describir la evidencia disponible sobre el uso de IA, incluyendo LLM y herramientas generativas, para la formación clínica práctica de estudiantes de medicina de pregrado. Esta revisión no busca estimar tamaños de efecto ni comparar formalmente intervenciones, sino caracterizar el panorama de aplicaciones, contextos, outcomes reportados y vacíos de conocimiento, lo cual es consistente con el objetivo y la utilidad de una revisión de alcance o *scoping review*.

2. Metodología

Esta revisión se desarrolló siguiendo el marco metodológico de Arksey y O'Malley (7). El reporte se realizó de acuerdo con la guía PRISMA-ScR, para asegurar transparencia y reproducibilidad.

Etapa 1: Identificación de la pregunta de investigación

Como primera etapa de esta revisión, se definió la siguiente pregunta de investigación: "¿Cómo se ha utilizado la inteligencia artificial en la formación clínica práctica de estudiantes de medicina de pregrado, en qué contextos y dominios competenciales, con qué resultados reportados y qué brechas metodológicas y éticas persisten?"

Etapa 2: Identificación de la literatura relevante

La búsqueda bibliográfica se realizó el 28 de enero de 2026 en PubMed/MEDLINE, Scopus y Web of Science (Core Collection). Se utilizaron operadores booleanos y una combinación de

vocabulario controlado y términos de texto libre relacionados con inteligencia artificial, educación médica y formación en habilidades clínicas. En PubMed/MEDLINE se emplearon términos MeSH como "Artificial Intelligence", "Machine Learning", "Deep Learning", "Students, Medical", "Education, Medical, Undergraduate" y "Clinical Competence"; además, en las tres bases se incorporaron términos de texto libre como *"large language model", "generative AI", "ChatGPT", *"chatbot", *"medical student", "clinical reasoning", *"clinical skill", "simulation", "objective structured clinical examination" y "mini-CEX". La estrategia completa de búsqueda para cada base de datos se presenta en el Anexo 1. Además, se utilizaron términos de exclusión para reducir estudios centrados en rendimiento algorítmico puro, como "model performance", "algorithm performance", "predictive model" o "diagnostic model".

Para la selección de estudios se definieron criterios de inclusión y exclusión resumidos en la tabla 1 del anexo. Se incluyeron estudios empíricos (cuantitativos, cualitativos o de métodos mixtos) publicados desde 2021 en adelante, en inglés, español o portugués, que evaluaran intervenciones educativas basadas en IA definida explícitamente (p. ej., ML/DL, LLM, IA generativa, NLP, visión computacional o algoritmos entrenados) y utilizadas con fines de enseñanza, entrenamiento, evaluación o retroalimentación en competencias clínicas. Los estudios debían involucrar estudiantes de medicina de pregrado (MD/MBBS o equivalente); en estudios con poblaciones mixtas, se incluyeron únicamente aquellos que reportaran resultados separables para pregrado.

Para efectos de esta revisión, se consideraron estudios que incluyeran desde la práctica clínica real supervisada hasta diversos contextos estructurados de simulación, tales como OSCE, Mini-CEX, laboratorios de destrezas, pacientes virtuales o digitales y casos simulados interactivos. Estos contextos fueron incluidos por compartir un foco en el desempeño de competencias clínicas observables y en la aplicación de habilidades en situaciones clínicas. Sin embargo, es relevante destacar que no se consideran equivalentes entre sí, ya que difieren en su grado de autenticidad pedagógica, complejidad contextual y potencial de transferencia al desempeño con pacientes reales. Esta distinción se incorporó desde el diseño de la revisión y se mantuvo durante el análisis, permitiendo interpretar la evidencia considerando dicha heterogeneidad.

Para evitar ambigüedades terminológicas, en esta revisión el término "formación clínica práctica" se utilizó en sentido amplio para referirse al continuo de escenarios orientados al desarrollo de competencias clínicas aplicadas, que incluye tanto la práctica clínica real supervisada como contextos estructurados de simulación, tales como OSCE, laboratorios de destrezas, pacientes virtuales o digitales y casos simulados interactivos. El término "simulación" se reservó para escenarios que no involucran pacientes reales, mientras que "contextos estructurados" se empleó para referirse a escenarios con criterios de desempeño explícitos y condiciones de evaluación estandarizadas, independientemente de si ocurren en simulación o en práctica clínica real. Se excluyeron estudios centrados exclusivamente en residentes, especialistas u otros profesionales de la salud, así como aquellos que describieran intervenciones educativas sin componente de IA o donde la IA se utilizara con fines no educativos. También, se excluyeron estudios que evaluaran únicamente satisfacción, percepción o actitudes hacia la IA sin resultados relacionados con aprendizaje/desempeño/implementación, y publicaciones no empíricas, incluyendo editoriales, comentarios, cartas al editor, ensayos de opinión, protocolos sin resultados y revisiones secundarias.

Etapa 3: Selección de estudios apropiados (screening)

Los registros identificados fueron importados a Rayyan para la detección de duplicación y el proceso de selección de forma manual. La elegibilidad se evaluó en dos fases (título/resumen y texto completo) utilizando los criterios de inclusión y exclusión predefinidos (tabla 1 anexo). El proceso fue realizado por cuatro revisores. Las discrepancias se resolvieron mediante discusión y consenso del equipo; cuando fue necesario, se efectuó una revisión adicional para adjudicación.

Etapa 4: Extracción, mapeo y gráficos de datos

Se diseñó un formulario de extracción (*charting form*) para registrar, de manera estandarizada, las características metodológicas y los hallazgos principales de los estudios incluidos. Las variables extraídas incluyeron: autor/año, país, diseño del estudio, tipo y número de participantes (incluyendo el nivel formativo cuando estuvo disponible), contexto educativo (clínica real supervisada y/o simulación), tecnología de IA (p. ej., LLM, ML/DL, NLP/visión), finalidad educativa (enseñanza, entrenamiento, evaluación o *feedback*), *outcomes* evaluados y resultados principales. Los datos extraídos se sintetizaron mediante dos productos complementarios. En primer lugar, se elaboró una tabla descriptiva (tabla 2 anexo) que resume, para cada estudio incluido, el país, diseño, participantes, tecnología de IA y principales resultados, con el objetivo de mapear el alcance de la evidencia y la heterogeneidad de enfoques, contextos y aplicaciones. En segundo lugar, se construyó una tabla comparativa (tabla 3 anexo) centrada en el subconjunto de estudios que reportaron comparaciones directas entre una intervención basada en IA y un comparador (p. ej., método tradicional, instrucción humana, cohorte histórica u otra condición), consignando la competencia evaluada, el comparador y el resultado principal reportado. Esta tabla permitió identificar en qué competencias y bajo qué diseños existe evidencia comparativa.

Adicionalmente, se incorporó a la tabla 2 una variable analítica derivada ("dominio de competencia") para clasificar cada estudio según la competencia clínica práctica predominante abordada por la intervención. Para ello, se utilizó una taxonomía cerrada de seis dominios: (1) entrevista clínica, historia clínica y comunicación; (2) razonamiento clínico y toma de decisiones diagnósticas; (3) imagenología e interpretación diagnóstica; (4) habilidades técnicas/procedimentales; (5) evaluación y feedback automatizado del desempeño (assessment-centric); y (6) competencia clínica integrada (p. ej., OSCE/Mini-CEX). Esta taxonomía fue construida como una variable analítica derivada, con fines de síntesis, a partir de la lectura de los estudios incluidos y del interés de la revisión por caracterizar la competencia clínica práctica predominante abordada por cada intervención. Su aplicación se llevó a cabo mediante una revisión basada en el consenso entre los revisores. Cuando surgieron discrepancias en la asignación del dominio primario, estas se resolvieron mediante discusión del equipo hasta lograr acuerdo. La asignación se realizó de manera mutuamente excluyente, definiendo un dominio primario por estudio. Para resolver ambigüedades, se aplicó la siguiente regla: si el outcome principal correspondía a consistencia, estabilidad o validez de la evaluación (p. ej., concordancia humano-IA o concordancia interevaluador), el estudio se clasificó en el dominio 5; si el outcome principal correspondía a mejora del desempeño del estudiante (pre/post o comparación con control), se clasificó en los dominios 1–4 o 6, según la competencia predominante.

A partir de esta clasificación, se elaboró un gráfico de barras que muestra el número de estudios por dominio de competencia (figura 2), así como un mapa geográfico que muestra la distribución de los países con publicaciones sobre el tema, utilizando una mayor intensidad de color para representar aquellos con mayor frecuencia de aparición en la muestra (figura 3). Finalmente, se muestra también un mapa de calor que representa la distribución de los estudios según dominio de competencia y tipo de IA (figura 4).

Etapa 5: Resumen y presentación de los resultados

Los resultados se sintetizaron mediante un enfoque descriptivo y narrativo. Se reportaron frecuencias y proporciones para caracterizar la distribución de estudios según país, diseño, tipo de participantes, contexto educativo (clínica real vs simulación), tipo de tecnología de IA y finalidad educativa. La evidencia se presentó agrupada por el "dominio de competencia" definido previamente y, cuando fue pertinente, por tipo de IA y contexto de implementación. El proceso de selección de estudios se documentó mediante un diagrama de flujo PRISMA-ScR, incluyendo el número de registros identificados, duplicados removidos, estudios evaluados a texto completo y razones de exclusión. Los hallazgos se presentaron en una tabla descriptiva (tabla 2 anexo) para

mapear el alcance y heterogeneidad de la evidencia, y en una tabla comparativa (tabla 3 anexo) para el subconjunto de estudios con comparación directa entre intervenciones basadas en IA y un comparador. Dada la heterogeneidad de diseños, intervenciones y outcomes, no se realizó síntesis cuantitativa ni meta-análisis, ni se interpretaron los resultados como una estimación global de efectividad. Los hallazgos comparativos se reportaron de manera descriptiva, destacando patrones, brechas metodológicas y áreas prioritarias para investigación futura.

Por otro lado, de acuerdo con las recomendaciones PRISMA-ScR, se señala explícitamente que la decisión de no realizar evaluación formal de calidad metodológica es consistente con el propósito de mapeo de una scoping review, pero implica que los hallazgos no permiten jerarquizar la solidez de la evidencia ni establecer recomendaciones basadas en la calidad de los estudios. Esta limitación afecta directamente la aplicabilidad de los resultados: los patrones comparativos reportados deben entenderse como señales descriptivas del campo, y no como base para recomendaciones de implementación clínica o curricular. Investigaciones futuras de tipo revisión sistemática deberían incorporar evaluación de riesgo de sesgo para avanzar hacia conclusiones sobre efectividad.

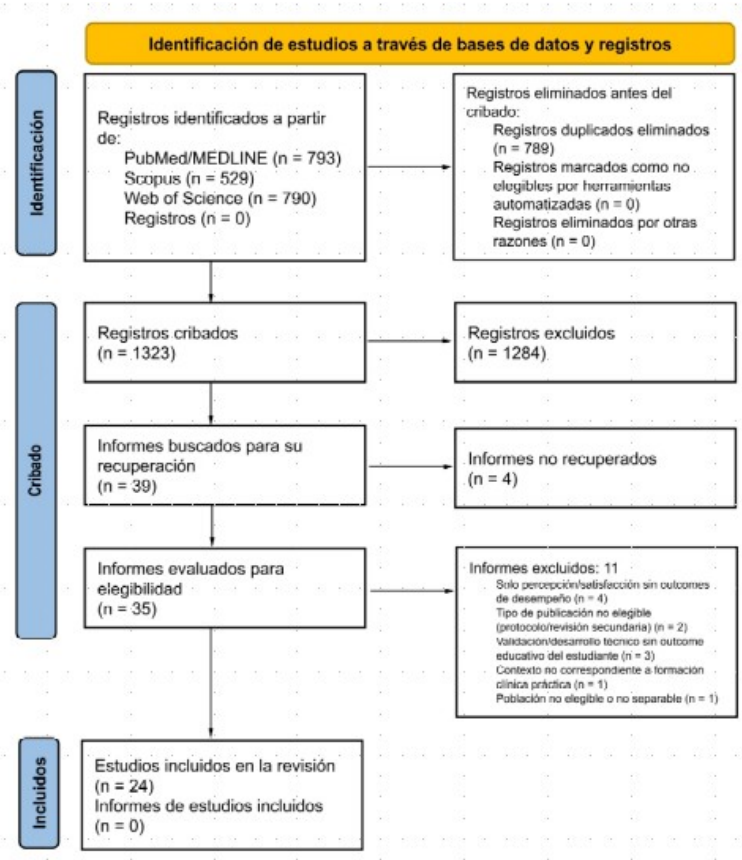


Figura 1. Diagrama de flujo PRISMA del proceso.

3. Resultados

3.1 Proceso de selección de resultados

El proceso de selección de estudios se presenta en la figura 1, de acuerdo con el diagrama PRISMA. En total, se identificaron 2112 registros, de los cuales 789 correspondían a duplicados, por lo que se incluyeron 1323 registros para la etapa de cribado. Estos fueron evaluados mediante revisión de título y resumen, aplicando los criterios de inclusión y exclusión definidos previamente en la sección de metodología. Tras esta etapa, se excluyeron 1284 registros, quedando 39 artículos

para revisión a texto completo. Durante el proceso de recuperación, 4 artículos no pudieron obtenerse debido a restricciones de acceso por pago. En consecuencia, se evaluaron 35 artículos en texto completo. De estos, 11 fueron excluidos por las siguientes razones: evaluar únicamente percepción/satisfacción (n = 4), corresponder a un tipo de publicación no elegible (n = 2), constituir validación o desarrollo técnico sin outcome educativo del estudiante (n = 3), presentar un contexto no correspondiente a formación clínica práctica (n = 1), o incluir una población no elegible o no separable (n = 1). Finalmente, se incluyeron 24 estudios en la revisión.

3.2 Características Generales de los Artículos Seleccionados

En total, se incluyeron 24 estudios que examinaron el uso de distintas tecnologías de IA en la formación clínica práctica de estudiantes de medicina. Los artículos fueron publicados entre 2022 y 2026, con una concentración en 2024 y 2025, lo que refleja un aumento reciente de publicaciones sobre aplicaciones educativas de IA, particularmente en contextos clínicos simulados y en intervenciones basadas en modelos generativos y modelos de lenguaje de gran escala (8-31).

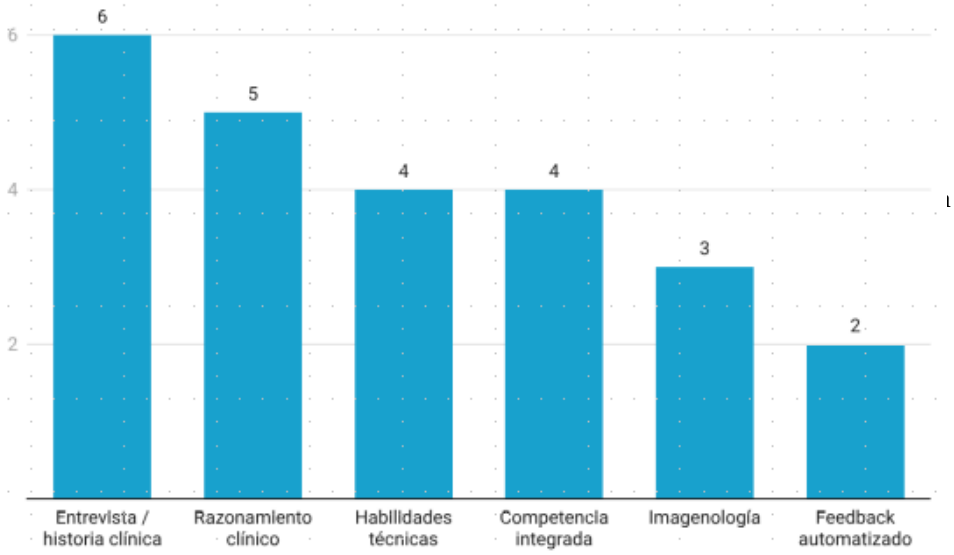


Figura 2. Distribución de artículos según competencia clínica principal.

Desde el punto de vista geográfico, los estudios se concentraron principalmente en China, Alemania y Canadá, con representación adicional de Japón, Suiza, Turquía, Hong Kong, Singapur y Estados Unidos, como se muestra en la figura 3 y en la tabla 2 del anexo. Esta distribución sugiere que la evidencia disponible proviene mayoritariamente de contextos con alta capacidad tecnológica e infraestructura digital para integrar IA en entornos de formación médica (10, 13-18, 24, 28-30).

En relación con el diseño metodológico, se identificó una combinación de ensayos clínicos aleatorizados, estudios controlados no aleatorizados, diseños prospectivos paralelos, estudios piloto o de factibilidad, estudios mixtos y trabajos centrados en la validación de sistemas de evaluación automatizada. En términos generales, una proporción importante de la muestra correspondió a estudios comparativos o cuasi-experimentales, especialmente en simulación clínica, entrevista médica, Mini-CEX y entrenamiento procedimental (8, 9, 11-13, 15, 17, 28, 31).

El tamaño muestral fue heterogéneo, desde estudios con muestras pequeñas propias de intervenciones piloto o ensayos en simulación hasta cohortes más amplias implementadas en contextos curriculares o institucionales (8, 9, 11, 12, 15, 27, 28). La mayoría de los participantes correspondió a estudiantes de medicina de pregrado, incluyendo estudiantes en etapas iniciales, intermedias o avanzadas de formación, y en algunos estudios se incorporaron además expertos

clínicos como referencia para validar criterios de evaluación, concordancia o calidad del feedback automatizado (10, 18, 31).

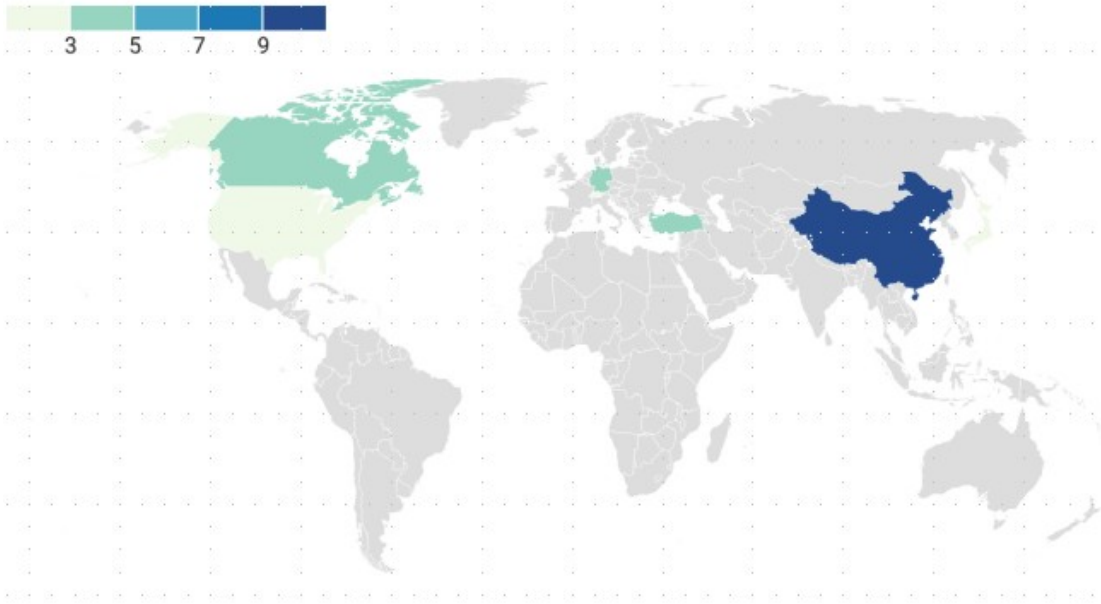


Figura 3. Representación geográfica de los artículos seleccionados.

En cuanto al contexto educativo, predominó el desarrollo de intervenciones en entornos simulados o estructurados, como pacientes simulados o virtuales, estaciones tipo OSCE, evaluaciones Mini-CEX, plataformas de tutoría virtual, laboratorios de habilidades y simulación quirúrgica (8, 9, 11-13, 15, 17, 20, 23, 25). En menor proporción, algunos estudios se integraron a contextos curriculares más cercanos a la práctica clínica o a flujos de trabajo diagnósticos reales, como cursos clínicos o sistemas PACS con soporte de IA (29-30).

Respecto a las tecnologías empleadas, los LLM y herramientas generativas constituyeron la categoría más frecuente, especialmente en intervenciones orientadas a entrevista clínica, razonamiento diagnóstico, pacientes simulados conversacionales y feedback automatizado (9, 10, 12, 16-18, 22-25, 27). También, se identificaron sistemas de IA para entrenamiento técnico/procedimental y simulación quirúrgica, incluyendo modelos de visión computacional, análisis multimodal y tutores adaptativos (8, 20, 28, 31), así como herramientas asociadas a interpretación diagnóstica e imagenología (21, 26, 30).

Finalmente, los outcomes reportados se concentraron principalmente en medidas de desempeño en escenarios simulados, puntuaciones estructuradas de competencia clínica (p. ej., OSCE, Mini-CEX), medidas de desempeño técnico y concordancia entre IA y evaluadores humanos, mientras que fueron menos frecuentes los estudios con seguimiento longitudinal o con evaluación de transferencia al entorno clínico real (10, 12, 13, 15, 17, 18, 20, 31).

3.3 Distribución según dominio de competencia

Con el fin de caracterizar en qué competencias clínicas prácticas se está utilizando la IA, los estudios se clasificaron en una taxonomía de seis dominios, cuya distribución se presenta en las figuras 1 y 2. Los dominios con mayor frecuencia fueron entrevista clínica, historia clínica y comunicación (n = 6) y razonamiento clínico y toma de decisiones diagnósticas (n = 5). El primer grupo incluyó estudios centrados en anamnesis, interacción clínico-paciente y práctica de entrevista mediante pacientes simulados conversacionales, chatbots o pacientes digitales (9, 11, 12, 16, 22, 25).

El segundo reunió intervenciones en las que la IA se utilizó para guiar razonamiento clínico, resolución de casos o toma de decisiones diagnósticas (14, 19, 24, 27, 29).

En una frecuencia intermedia se ubicaron los dominios de habilidades técnicas/procedimentales (n = 4) y competencia clínica integrada (n = 4). En el primero se agruparon estudios sobre entrenamiento técnico o quirúrgico con simulación, tutorización adaptativa y feedback automatizado del desempeño psicomotor (8, 20, 28, 31). En el segundo se incluyeron intervenciones evaluadas mediante medidas globales de competencia clínica, como Mini-CEX o desempeño integrado en escenarios clínicos estructurados (13, 15, 17, 23).

Los dominios menos representados fueron imagenología e interpretación diagnóstica (n = 3) y evaluación y feedback automatizado del desempeño (n = 2). En imagenología se identificaron estudios que integraron IA a la interpretación diagnóstica o al flujo de trabajo con imágenes, incluyendo entrenamiento en ultrasonido, queratitis e integración de IA en PACS (21, 26, 30). Por su parte, el dominio de feedback automatizado incluyó estudios centrados principalmente en consistencia, estabilidad o concordancia entre IA y evaluadores humanos en tareas de feedback o scoring (10, 18).

En conjunto, esta distribución muestra que la evidencia se concentra principalmente en dominios que pueden ser simulados, estructurados y escalados con relativa facilidad mediante IA, especialmente entrevista clínica, razonamiento basado en casos y entrenamiento técnico en simulación (9, 14, 16, 20, 28).

3.4 Relación entre dominio de competencia y tipo de IA

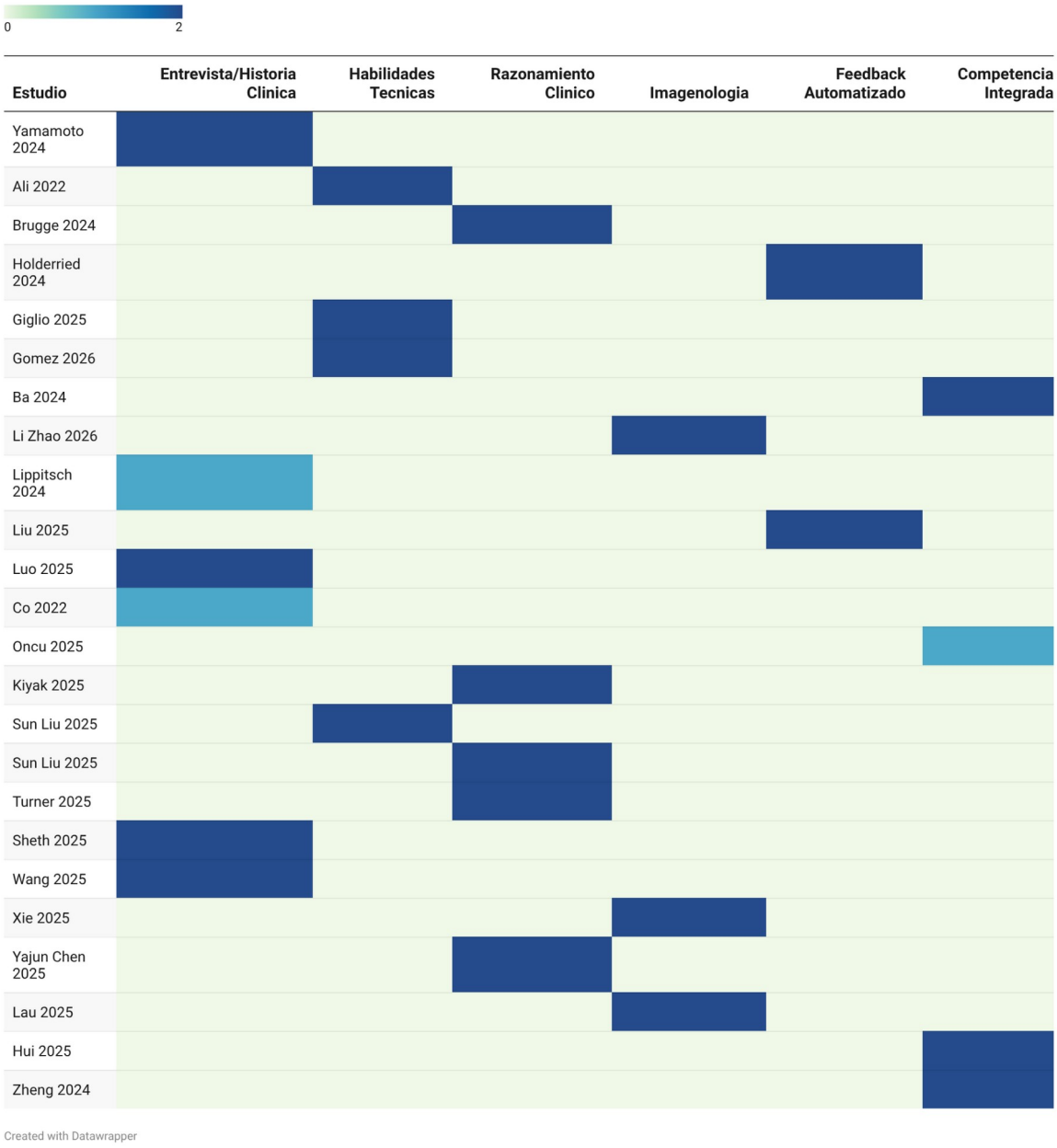
La relación entre dominio de competencia y tipo de IA se representa en la figura 2. En términos generales, los LLM fueron la tecnología predominante en los dominios de entrevista clínica/comunicación y razonamiento clínico, donde se utilizaron para construir pacientes simulados conversacionales, apoyar anamnesis, guiar discusiones clínicas y ofrecer feedback o apoyo a la toma de decisiones (9, 10, 12, 14, 16, 19, 22, 24, 25, 28).

En contraste, en el dominio de habilidades técnicas/procedimentales se observaron con mayor frecuencia sistemas de IA combinados con simulación, visión computacional, análisis de movimiento o tutorización adaptativa, orientados a monitorizar desempeño, identificar errores y proporcionar retroalimentación estructurada en tiempo real (8, 20, 28, 31).

En imagenología e interpretación diagnóstica, las intervenciones se asociaron más bien a herramientas integradas a entornos diagnósticos o a materiales de entrenamiento asistidos por IA, lo que incluyó sistemas PACS con soporte de IA, generación de imágenes y apoyo a tareas interpretativas específicas (21, 26, 30). Por otra parte, en el dominio feedback automatizado, la IA se utilizó principalmente para estandarizar evaluación y feedback, con énfasis en consistencia o concordancia frente a evaluadores humanos (10, 18).

En conjunto, el patrón general sugiere que los LLM dominan las aplicaciones centradas en interacción verbal, razonamiento y pacientes simulados, mientras que otros enfoques de ML/DL, visión computacional o sistemas multimodales aparecen con mayor frecuencia en dominios técnicos o de evaluación estructurada (8, 10, 18, 20, 31).

Figura 4. Mapa de calor de la distribución de los estudios por dominio de competencia clínica.



3.5 Subconjunto de estudios con comparación directa

La tabla 3 del anexo resume el subconjunto de estudios que incluyeron comparación directa entre una intervención basada en IA y un comparador. Este subconjunto estuvo conformado por estudios que compararon IA con métodos tradicionales de enseñanza, instrucción humana, cohortes históricas u otras condiciones de intervención, y constituyó una parte relevante del corpus total (8, 9, 11-13, 15-17, 20, 25, 28, 31).

En el dominio de entrevista clínica y comunicación, las comparaciones directas se centraron en pacientes simulados, chatbots o pacientes digitales para entrenamiento de anamnesis, mostrando resultados que variaron entre desempeño comparable y mejoras en algunas medidas específicas de rendimiento o estructura de entrevista (9, 11, 12, 16, 25). En competencia clínica integrada, se observaron comparaciones en contextos evaluados mediante Mini-CEX u otros formatos estructurados, comparando modalidades tradicionales con apoyo de IA o LLM (13, 15, 17).

En habilidades procedimentales, los estudios compararon tutorización o feedback basado en IA con instrucción experta o enseñanza tradicional, utilizando escalas de desempeño técnico en simulación como principal outcome (8, 20, 28, 31). En conjunto, este subconjunto muestra que existe evidencia comparativa en múltiples dominios de competencia. No obstante, los estudios difieren sustancialmente en sus comparadores, outcomes, duración de seguimiento y contextos de implementación, por lo que estos hallazgos deben interpretarse como un mapeo descriptivo de patrones comparativos reportados, más que como una base para inferencias agregadas sobre efectividad relativa (8, 9, 12, 13, 28).

3.6 Síntesis narrativa de patrones y vacíos

En términos generales, la evidencia reciente se concentró en tres ejes principales: (i) entrenamiento de entrevista clínica y anamnesis mediante pacientes simulados o chatbots, (ii) apoyo al razonamiento clínico en escenarios basados en casos, y (iii) entrenamiento y evaluación de habilidades técnicas/procedimentales en simulación (8-10, 12, 14, 16, 20, 24, 28, 31).

Por el contrario, se observó una menor densidad relativa de estudios en evaluación automatizada a gran escala, en integración sostenida dentro de contextos clínicos reales, y en investigaciones que evaluaran la retención longitudinal de competencias o su transferencia al desempeño clínico con pacientes reales (10, 18, 30). Asimismo, aunque varios estudios utilizaron medidas estructuradas de desempeño y algunos incluyeron comparadores directos, la evidencia continuó siendo heterogénea en términos de outcomes, duración de seguimiento, comparadores y escenarios de implementación (8, 9, 12, 13, 15, 17).

4. Discusión

La presente revisión de alcance evidencia que la incorporación de IA en la formación clínica práctica de estudiantes de medicina de pregrado se encuentra en una fase de expansión acelerada, pero aún en consolidación conceptual, metodológica y pedagógica. Más que demostrar de manera concluyente su superioridad sobre los métodos tradicionales, los hallazgos sugieren que la IA está redefiniendo las condiciones bajo las cuales se adquieren, practican y evalúan las competencias clínicas, particularmente en contextos donde la escalabilidad y la estandarización son limitantes estructurales.

Uno de los elementos más relevantes fue la concentración de la evidencia en dominios altamente estructurables y simulables, como la entrevista clínica, el razonamiento diagnóstico basado en casos y las habilidades procedimentales en entornos controlados, tal como se observa en la Figura 1 y en la tabla 2. Este patrón es consistente con revisiones previas de mayor amplitud, que muestran que las aplicaciones de la IA en educación médica tienden a concentrarse en áreas como el aprendizaje interactivo, la evaluación de habilidades quirúrgicas, y la interpretación diagnóstica, es decir, en contextos donde las tareas pueden estandarizarse, simularse y repetirse con relativa facilidad (3, 5). En particular, la literatura reciente sobre pacientes basados en modelos LLM sugiere que estos sistemas muestran especial promesa para el entrenamiento en habilidades comunicacionales y simulaciones clínicas conversacionales, precisamente porque operan bien en tareas verbalizables, con escenarios relativamente acotados y posibilidad de retroalimentación inmediata (32). En este sentido, más que reemplazar la enseñanza clínica tradicional, la IA parece estar optimizando aquellos componentes del aprendizaje que son más susceptibles de estandarización, lo que podría favorecer una práctica deliberada más frecuente, accesible y potencialmente individualizada para cada estudiante. Sin embargo, esta misma afinidad con dominios altamente simulables también ayuda a explicar por qué la mayor parte de la evidencia sigue concentrándose en contextos estructurados y simulados, más que en escenarios clínicos reales o longitudinales (32). Esta distribución debe leerse, además, considerando que la formación clínica

práctica fue conceptualizada en esta revisión como un continuo que abarca escenarios con distinto grado de autenticidad pedagógica, control del entorno y cercanía al desempeño clínico con pacientes reales.

Sumado a lo anterior, esta misma concentración de evidencia pone de relieve una limitación importante: la escasa evidencia en contextos clínicos reales y en competencias más complejas, dinámicas e interpersonales, que dependen de la incertidumbre, la variabilidad del paciente y la toma de decisiones contextualizada. Esta observación es coherente con revisiones previas que muestran que gran parte de las aplicaciones educativas de IA se han desarrollado en áreas altamente estructuradas, tales como el aprendizaje interactivo, la evaluación procedimental o la interpretación diagnóstica, y con evidencia predominantemente generada en entornos simulados o controlados (3, 5). En la misma línea, revisiones recientes sobre outcomes educativos medibles señalan que la evidencia disponible sigue sustentándose principalmente en estudios unicéntricos, con muestras pequeñas y duración limitada, lo que dificulta establecer si los beneficios observados se traducen en retención longitudinal o en mejor desempeño clínico en estudios reales (6). De manera más específica, la literatura sobre pacientes virtuales basados en LLM sugiere que, aunque estos sistemas muestran potencial para el entrenamiento y para el desarrollo de habilidades comunicacionales, no reemplazan las interacciones del mundo real ni reproducen plenamente la complejidad relacional y contextual del encuentro clínico (32). En este sentido, la escasa evaluación de transferencia al entorno clínico-asistencial sugiere que, hasta ahora, buena parte de la evidencia permanece en un nivel intermedio entre el aprendizaje simulado y el desempeño clínico auténtico.

Desde un marco educativo que articula práctica deliberada, aprendizaje autorregulado y andamiaje instruccional, esta distribución sugiere que la IA parece adaptarse mejor, por ahora, a competencias que pueden descomponerse en tareas explícitas, observables y repetibles, donde los criterios de desempeño son relativamente estables y la retroalimentación puede entregarse de manera inmediata y estandarizada. Desde la lógica de la práctica deliberada, estas herramientas podrían ampliar oportunidades de repetición orientada a objetivos y feedback frecuente; desde el aprendizaje autorregulado, podrían apoyar el monitoreo del desempeño, la identificación de brechas y el ajuste progresivo del aprendizaje; y desde el andamiaje instruccional, podrían ofrecer apoyos graduados que faciliten la progresión hacia mayor autonomía (3, 32, 34). Esto ayuda a entender su mayor presencia en anamnesis estructurada, razonamiento basado en casos y entrenamiento procedimental en simulación, así como el interés reciente por pacientes virtuales conversacionales y experiencias de aprendizaje más personalizadas (3, 32, 34). En estos dominios, su principal aporte no parecería radicar en sustituir la docencia humana, sino en ampliar oportunidades de práctica deliberada, ofrecer andamiaje más constante y apoyar trayectorias de aprendizaje más accesibles y adaptativas. Sin embargo, esa misma lógica plantea un límite relevante: cuando la competencia clínica depende de ambigüedad, juicio contextual, negociación relacional o integración dinámica de múltiples señales del entorno, una excesiva estandarización corre el riesgo de simplificar indebidamente el aprendizaje y empobrecer la complejidad del encuentro clínico. En este sentido, más que concebir la IA como una solución homogénea para toda la formación clínica, los hallazgos sugieren interpretar su valor pedagógico de manera situada, según el tipo de competencia involucrada y el contexto formativo en que se implementa, manteniendo cautela frente a extrapolaciones amplias sobre su alcance educativo (3, 5, 6).

Por otro lado, un subconjunto de estudios comparativos reportó resultados comparables o, en algunos casos, favorables para intervenciones basadas en IA, particularmente en escenarios donde la IA se empleó como apoyo para feedback, tutorización o razonamiento clínico (8, 9, 17, 28, 31). No obstante, estos hallazgos deben interpretarse con cautela, ya que los estudios difirieron en comparadores, outcomes, condiciones de implementación e intensidad del apoyo educativo ofrecido, lo que limita la posibilidad de extraer conclusiones agregadas sobre efectividad relativa. En este subconjunto comparativo, varios estudios utilizaron comparadores correspondientes a

estrategias docentes con menor intensidad de retroalimentación, menor estandarización o menor disponibilidad de apoyo inmediato, lo que dificulta atribuir los beneficios observados exclusivamente al componente de IA de la intervención (8, 9, 17, 28, 31). Esta característica limita la capacidad de establecer relaciones causales entre el uso de IA y los resultados reportados, ya que las diferencias podrían reflejar, al menos en parte, desigualdades en la calidad, estructuración o intensidad de la intervención educativa más que un efecto específico del componente tecnológico. Esto es coherente con revisiones recientes que, si bien describen hallazgos prometedores en determinados dominios, también subrayan la heterogeneidad de comparadores, outcomes y diseños, así como la necesidad de estudios metodológicamente más robustos antes de plantear conclusiones sobre superioridad general o efectividad comparativa de la IA frente a estrategias docentes convencionales (5-6). En consecuencia, los resultados favorables reportados en este subconjunto deben interpretarse como señales comparativas preliminares y no como evidencia causal directa de superioridad de la IA, especialmente cuando los comparadores no controlan de manera equivalente la intensidad formativa.

Asimismo, emerge un cambio relevante en el rol pedagógico de la IA: desde herramienta de acceso a contenidos hacia un agente activo dentro del proceso educativo. En particular, los LLM han favorecido el desarrollo de pacientes simulados conversacionales y sistemas de retroalimentación automatizada que no sólo entregan información, sino que interactúan, responden y ajustan la experiencia de aprendizaje según el desempeño del estudiante (9, 10, 17, 22, 25, 32). Este desplazamiento es coherente con la literatura, que describe a los LLM como tecnologías con potencial para ofrecer experiencias de aprendizaje más personalizadas, retroalimentación inmediata y apoyo más continuo al aprendizaje autodirigido, especialmente en contextos de simulación y entrenamiento clínico estructurado (32, 34). En este sentido, la IA parece acercar la educación médica a modelos más centrados en el aprendizaje activo y personalizado, en sintonía con principios del aprendizaje autorregulado y la práctica deliberada. Sin embargo, este cambio también introduce desafíos importantes en términos de confiabilidad, transparencia, sesgos y control pedagógico, particularmente por la naturaleza probabilística y, en ocasiones, no determinista de estos sistemas, lo que refuerza la necesidad de integración curricular supervisada y de evaluación pedagógica rigurosa antes de asumir su beneficio sostenido (5, 6, 32, 33, 35).

Otro aspecto crítico fue la heterogeneidad metodológica observada en los estudios incluidos. La variabilidad en diseños, tamaños muestrales, outcomes y comparadores limita la posibilidad de establecer conclusiones robustas sobre la efectividad de estas intervenciones. Este hallazgo es coherente con revisiones previas que describen una literatura amplia, pero metodológicamente heterogénea, con predominio de estudios en dominios específicos y con comparadores difícilmente homologables entre sí (3, 5). En la misma línea, una revisión sistemática reciente sobre outcomes educativos medibles en educación en profesionales de la salud concluyó que la evidencia disponible sigue siendo globalmente pobre, basada sobre todo en estudios unicéntricos, con muestras pequeñas y centrada en resultados correspondientes a niveles intermedios de la jerarquía de Kirkpatrick, sin evaluación de comportamientos más complejos en el lugar de trabajo ni de impacto en resultados asistenciales (6). En este contexto, la predominancia de estudios de corto plazo dificulta interpretar el impacto real de estas intervenciones más allá del desempeño inmediato, lo cual es particularmente relevante en educación médica, donde los outcomes de interés incluyen también retención, transferencia a la práctica clínica y repercusión sobre la calidad de la atención a pacientes (5-6).

Esta heterogeneidad constituye, además, una limitación estructural de la presente revisión que merece ser destacada explícitamente. Al no haberse realizado evaluación formal de calidad metodológica ni de riesgo de sesgo, decisión consistente con el propósito exploratorio de una scoping review, no es posible jerarquizar la solidez de los estudios incluidos ni comparar la robustez relativa de sus hallazgos. En consecuencia, los resultados comparativos reportados deben

leerse estrictamente como señales descriptivas preliminares, especialmente en un corpus heterogéneo en diseños, comparadores, outcomes y contextos educativos. La ausencia de evaluación limita también la posibilidad de valorar la susceptibilidad a sesgos y de extraer inferencias sobre efectividad comparativa, por lo que cualquier lectura de los patrones reportados debe mantenerse en el nivel del mapeo descriptivo del campo.

En este marco, la decisión de excluir estudios centrados exclusivamente en satisfacción, percepción o actitudes hacia la IA respondió al interés de privilegiar evidencia con mayor pertinencia práctica para esta revisión, es decir, aquella vinculada a outcomes educativos o de implementación más sustantivos. Aunque las percepciones y la aceptabilidad son dimensiones relevantes en campos emergentes como lo es la IA, incluir de manera amplia estudios sólo perceptuales habría podido inflar artificialmente el volumen de evidencia disponible sin aportar información suficiente sobre aprendizaje, desempeño o integración educativa efectiva. No obstante, esta decisión también implica que aspectos como aceptabilidad, adopción o barreras percibidas de implementación quedaron representados de manera indirecta en el corpus final.

Desde una perspectiva sistémica, la distribución geográfica de los estudios sugiere que la implementación de IA en educación médica está fuertemente condicionada por la disponibilidad de infraestructura digital, capacidad tecnológica y recursos institucionales. Este patrón, observado también en otras revisiones del área, plantea el riesgo de ampliar brechas entre contextos educativos, favoreciendo de manera desproporcionada a instituciones con mayor acceso a plataformas, conectividad, soporte técnico y capacidad de integración curricular. En consecuencia, la discusión sobre IA en educación médica no puede desligarse de consideraciones de equidad, acceso, transparencia y gobernanza, especialmente cuando la evidencia proviene de forma predominante de entornos con alta disponibilidad tecnológica y cuando los marcos éticos recientes insisten en la necesidad de implementar estas herramientas bajo principios de justicia, seguridad, responsabilidad y supervisión pedagógica (3, 32).

La tabla 1 muestra una síntesis del estado actual del conocimiento: lo que sabemos y lo que no sabemos sobre el uso de IA en la formación clínica práctica de estudiantes de medicina de pregrado.

Tabla 1. Síntesis del estado actual del conocimiento: lo que sabemos y lo que no sabemos sobre el uso de IA en la formación clínica práctica de estudiantes de medicina de pregrado.

Dimensión	Qué sabemos	Qué no sabemos / Brechas identificadas
Tecnologías predominantes	Los LLM y herramientas generativas son las tecnologías más frecuentemente estudiadas, especialmente en entrevista clínica, razonamiento diagnóstico y pacientes simulados conversacionales	Existe escasa evidencia sobre aplicaciones de IA no-LLM en dominios complejos e integrados; la mayoría de las herramientas evaluadas son prototipos o versiones experimentales
Contextos de implementación	La evidencia se concentra en simulación estructurada (OSCE, Mini-CEX, pacientes virtuales, laboratorios de destrezas)	Existe evidencia muy limitada en práctica clínica real supervisada; se desconoce cómo se comportan estas herramientas fuera de entornos controlados
Resultados educativos inmediatos	Algunos estudios comparativos reportan resultados comparables o	No hay evidencia robusta sobre retención longitudinal de

	favorables a la IA en medidas de desempeño inmediato en simulación	competencias ni sobre transferencia al entorno clínico real con pacientes
Efectividad comparativa	Existe evidencia comparativa en múltiples dominios (entrevista clínica, habilidades técnicas, razonamiento clínico, competencia integrada)	Los comparadores son heterogéneos y frecuentemente menos intensivos en retroalimentación, lo que impide atribuir los efectos observados específicamente al componente tecnológico
Marco pedagógico y ético	Se reconoce la necesidad de marcos de gobernanza, supervisión curricular y principios éticos para la integración de IA	No existen marcos pedagógicos ni éticos formales validados; las implicaciones para el diseño curricular permanecen mayormente implícitas en la literatura
Equidad y contexto	La distribución geográfica muestra concentración en contextos con alta capacidad tecnológica e infraestructura digital	Ausencia de estudios en entornos de recursos limitados o contextos de baja y mediana renta; riesgo de ampliar brechas educativas estructurales

Desde una perspectiva de diseño curricular, los hallazgos sugieren que la integración de IA en formación clínica práctica requiere decisiones pedagógicas explícitas sobre el rol que se le asigna: si como sustituto de oportunidades de práctica escasas, como andamiaje para el aprendizaje autodirigido, o como herramienta de evaluación formativa complementaria. En cada caso, las implicaciones para el diseño instruccional, la supervisión docente y la evaluación de competencias difieren sustancialmente. Independientemente del modelo curricular —ya sea basado en competencias, en resultados o en etapas de formación progresiva—, la incorporación responsable de IA requerirá definir con claridad los objetivos pedagógicos que se busca alcanzar, los criterios de calidad de las herramientas adoptadas, y los mecanismos de supervisión y evaluación continua de su impacto educativo.

Finalmente, los hallazgos refuerzan la necesidad de avanzar hacia una agenda de investigación más madura, que trascienda estudios de factibilidad o eficacia en entornos controlados. En línea con revisiones recientes, futuras investigaciones deberían priorizar diseños longitudinales, evaluación en contextos clínicos reales, medición de transferencia de competencias y comparación con intervenciones pedagógicas optimizadas –y no únicamente con métodos tradicionales de menor intensidad formativa–. Asimismo, es deseable una mayor estandarización de outcomes, comparadores y criterios de reporte, de modo que la futura literatura permita distinguir con mayor claridad entre efectos atribuibles al componente tecnológico y aquellos derivados de mayor estructuración, frecuencia de retroalimentación o intensidad de apoyo educativo (5-6).

En conjunto, la evidencia disponible sugiere que la IA podría contribuir de manera relevante a la formación clínica práctica, aunque parte de estos beneficios podría estar mediada por una mayor estructuración e intensidad del apoyo educativo, más que por el componente tecnológico de forma aislada. En este contexto, la IA no debe entenderse necesariamente como un reemplazo de los modelos docentes existentes, sino como un recurso capaz de complementar y rediseñar ciertos componentes de la enseñanza. Su valor parece concentrarse, por ahora, en la posibilidad de escalar oportunidades de práctica, estandarizar la retroalimentación y favorecer experiencias de

aprendizaje más accesibles y, en algunos casos, más personalizadas. Sin embargo, su integración efectiva requerirá un enfoque crítico, pedagógicamente guiado y basado en evidencia, que considere no solo su promesa tecnológica sino también sus limitaciones metodológicas, éticas y contextuales (3, 5-6, 35).

5. Conclusiones

- Esta revisión de alcance permitió mapear y caracterizar la evidencia actual sobre el uso de IA en la formación clínica práctica de estudiantes de medicina de pregrado. Los hallazgos muestran que la IA se ha integrado principalmente en escenarios simulados y en tareas estructuradas, como la entrevista clínica, el razonamiento diagnóstico y el entrenamiento procedimental, con un predominio creciente de herramientas basadas en modelos de lenguaje de gran escala.
- En suma, la evidencia sugiere que estas tecnologías pueden ampliar las oportunidades de práctica deliberada, estandarizar la retroalimentación y apoyar procesos de aprendizaje más accesibles y, en algunos casos, más personalizados. No obstante, su implementación actual se concentra en etapas intermedias del aprendizaje, con limitada evidencia sobre transferencia al entorno clínico real, impacto longitudinal y desempeño en contextos complejos, no estructurados o altamente variables.
- Asimismo, la heterogeneidad de los diseños, outcomes y comparadores, junto con la escasez de evaluaciones en contextos auténticos y de marcos pedagógicos y éticos consistentes, pone de relieve la necesidad de avanzar hacia investigaciones metodológicamente más robustas y pedagógicamente informadas. En este contexto, futuras investigaciones deberían priorizar la evaluación de la IA en entornos clínicos reales, la medición de transferencia de competencias y el desarrollo de marcos éticos y educativos que orienten su integración responsable. En suma, la IA se perfila como una herramienta prometedora para complementar y rediseñar ciertos componentes de la formación clínica, aunque su adopción efectiva requerirá una implementación crítica, contextualizada y basada en evidencia.

Financiación: No ha habido financiación.

Declaración de conflicto de interés: Los autores declaran no tener ningún conflicto de interés.

Declaración de IA generativa y tecnologías asistidas por IA en el proceso de redacción: Durante la elaboración del siguiente trabajo, se ha utilizado Inteligencia Artificial de forma limitada para mejorar la redacción. Posteriormente, el contenido ha sido revisado y editado, asumiendo total responsabilidad por el contenido de la publicación.

Contribuciones de los autores: J.B. participó en la búsqueda bibliográfica, revisión de la literatura, redacción, y revisión del manuscrito, elaboración de figuras e imágenes. D.P. y C.M. participaron en la búsqueda bibliográfica, revisión de la literatura, redacción y revisión del manuscrito. T.C. participó en la búsqueda bibliográfica, revisión de la literatura, redacción y revisión del manuscrito, así como en la elaboración de figuras e imágenes. A.H. participó en la revisión y edición del manuscrito como experto en educación médica. Todos los autores leyeron y aprobaron la versión final del manuscrito.

6. Referencias

1. Dewan P, Khalil S, Gupta P. Objective structured clinical examination for teaching and assessment: Evidence-based critique. *Clin Epidemiol Glob Health.* **2024**, 25, 101477. https://doi.org/10.1016/j.cegh.2023.101477
2. Tozsin A, Ucmak H, Soyturk S, Aydin A, Gozen AS, Al Fahim M, Güven S, Ahmed K. The role of artificial intelligence in medical education: A systematic review. *Surg Innov.* **2024**, 31(4), 415-423. https://doi.org/10.1177/15533506241248239

3. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, Hanson J, Haas M, Spadafore M, Grafton-Clarke C, Gasiea RY, Michie C, Corral J, Kwan B, Dolmans D, Thammasitboon S. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach.* **2024**, 46(4), 446-470. https://doi.org/10.1080/0142159x.2024.2314198
4. Aster A, Laupichler MC, Rockwell-Kollmann T, Masala G, Bala E, Raupach T. ChatGPT and other large language models in medical education - Scoping literature review. *Med Sci Educ.* **2024**, 35(1), 555-567. https://doi.org/10.1007/s40670-024-02206-6
5. Shaw K, Henning MA, Webster CS. Artificial intelligence in medical education: a scoping review of the evidence for efficacy and future directions. *Med Sci Educ.* **2025**, 35(3), 1803-1816. https://doi.org/10.1007/s40670-025-02373-0
6. Feigerlova E, Hani H, Hothersall-Davies E. A systematic review of the impact of artificial intelligence on educational outcomes in health professions education. *BMC Med Educ.* **2025**, 25, 129. https://doi.org/10.1186/s12909-025-06719-5
7. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol.* **2005**, 8(1), 19-32. https://doi.org/10.1080/1364557032000119616
8. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. 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. https://doi.org/10.1001/jamanetworkopen.2021.49008
9. Co M, Yuen THJ, Cheung HH. Using clinical history taking chatbot mobile app for clinical bedside teachings – A prospective case control study. *Heliyon.* **2022**, 8(6), e09751. https://doi.org/10.1016/j.heliyon.2022.e09751
10. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, et al. A language model-powered simulated patient with automated feedback for history taking: Prospective study. *JMIR Med Educ.* **2024**, 10, e59213. https://doi.org/10.2196/59213
11. Lippitsch A, Steglich J, Ludwig C, Kellner J, Hempel L, Stoevesandt D, et al. Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars. *Comput Methods Programs Biomed.* **2024**, 244, 107964. https://doi.org/10.1016/j.cmpb.2023.107964
12. 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. https://doi.org/10.2196/58753
13. Ba H, Zhang L, Yi Z. Enhancing clinical skills in pediatric trainees: A comparative study of ChatGPT-assisted and traditional teaching methods. *BMC Med Educ.* **2024**, 24(1), 558. https://doi.org/10.1186/s12909-024-05565-1
14. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Geist M, Jeselsohn M, et al. Large language models improve clinical decision making of medical students through patient simulation and structured feedback: A randomized controlled trial. *BMC Med Educ.* **2024**, 24(1), 1391. https://doi.org/10.1186/s12909-024-06399-7
15. Zheng K, Shen Z, Chen Z, et al. Application of AI-empowered scenario-based simulation teaching mode in cardiovascular disease education. *BMC Med Educ.* **2024**, 24(1), 1003. https://doi.org/10.1186/s12909-024-05977-z.
16. Luo MJ, Bi S, Pang J, Liu L, Tsui CK, Lai Y, et al. A large language model digital patient system enhances ophthalmology history taking skills. *NPJ Digit Med.* **2025**, 8(1), 502. https://doi.org/10.1038/s41746-025-01841-6.
17. Hui Z, Zewu Z, Jiao H, et al. Application of ChatGPT-assisted problem-based learning teaching method in clinical medical education. *BMC Med Educ.* **2025**, 25, 50. https://doi.org/10.1186/s12909-024-06321-1.
18. Liu Y, Shi C, Wu L, Lin X, Chen X, Zhu Y, et al. Development and validation of a large language model-based system for medical history-taking training: Prospective multicase

study on evaluation stability, human-AI consistency, and transparency. *JMIR Med Educ.* **2025**, 11, e73419. https://doi.org/10.2196/73419.

19. Kıyak YS, Emekli E, İş Kara T, Coşkun Ö, Budakoğlu İİ. AI teaches surgical diagnostic reasoning to medical students: Evidence from an experiment using a fully automated, low-cost feedback system. *J Surg Educ.* **2025**, 82(10), 103639. https://doi.org/10.1016/j.jsurg.2025.103639
20. Giglio B, Albeloushi A, Alhaj AK, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial intelligence-augmented human instruction and surgical simulation performance: A randomized clinical trial. *JAMA Surg.* **2025**, 160(9), 993-1003. https://doi.org/10.1001/jamasurg.2025.2564.
21. Lau YH, Acharyya S, Wee CWL, Xu H, Pulido Saclolo R, Cao K, et al. Effectiveness of traditional, artificial intelligence-assisted, and virtual reality training modalities for focused cardiac ultrasound skill acquisition: A randomised controlled study. *Ultrasound J.* **2025**, 17(1), 61. https://doi.org/10.1186/s13089-025-00469-7.
22. Sheth U, Lo M, McCarthy J, Baath N, Last N, Guo E, et al. Understanding the role of large language model virtual patients in developing communication and clinical skills in undergraduate medical education. *Int Med Educ.* **2025**, 4, 39. https://doi.org/10.3390/ime4040039
23. Öncü S, Torun F, Ülkü HH. AI-powered standardised patients: Evaluating ChatGPT-4o's impact on clinical case management in intern physicians. *BMC Med Educ.* **2025**, 25(1), 278. https://doi.org/10.1186/s12909-025-06877-6.
24. Turner L, Kelleher M, Overla S, Zheng W, Gregath A, Gharib M, et al. Harnessing the generative power of AI to move closer to personalized medical education. *Acad Med.* **2025**, 100(12), 1447-1451. https://doi.org/10.1097/ACM.0000000000006185.
25. Wang Z, Fan TT, Li ML, Zhu NJ, Wang XC. Feasibility study of using GPT for history-taking training in medical education: A randomized clinical trial. *BMC Med Educ.* **2025**, 25(1), 1030. https://doi.org/10.1186/s12909-025-07614-9.
26. Xie W, Yuan Z, Si Y, Huang Z, Li Y, Wu F, et al. Enhancing medical students' diagnostic accuracy of infectious keratitis with AI-generated images. *BMC Med Educ.* **2025**, 25(1), 1027. https://doi.org/10.1186/s12909-025-07592-y
27. Chen Y. Evaluation of the impact of AI-driven personalized learning platform on medical students' learning performance. *Front Med (Lausanne).* **2025**, 12, 1610012. https://doi.org/10.3389/fmed.2025.1610012
28. Sun Y, Liu F. Evaluating the impact of AI-tutoring versus expert human instruction on surgical skills in medical students. *Educ Inf Technol.* **2025**, 30(18), 26413-26431. https://doi.org/10.1007/s10639-025-13779-z.
29. Sun Y, Liu F. Real-world implementation of an AI learning tool-MetaGP-Edu in medical education: A multi-center cohort study. *Comput Educ.* **2025**, 237, 105388. https://doi.org/10.1016/j.compedu.2025.105388.
30. Li J, Zhao H. Workflow-embedded AI as a cognitive scaffold: A randomized trial on knowledge retention and diagnostic competency in undergraduate radiology education. *Eur J Radiol Open.* **2026**, 16, 100724. https://doi.org/10.1016/j.ejro.2026.100724.
31. Gomez C, Seenivasan L, Zou X, Yoon J, Chu S, Leong A, et al. Explainable AI for automated user-specific feedback in surgical skill acquisition. In: Guo X, Jin Y, Lamdouar H, Ouyang C, Men Q, Sahu M, Vedula SS, editors. Human-AI collaboration—First international workshop, HAIC 2025, held in conjunction with MICCAI 2025, proceedings. Lecture Notes in Computer Science, vol. 16214. Cham: Springer; **2026**, 25-34. https://doi.org/10.1007/978-3-032-08970-0_3.
32. Zeng J, Qi W, Shen S, Liu X, Li S, Wang B, et al. Embracing the future of medical education with large language model-based virtual patients: Scoping review. *J Med Internet Res.* **2025**, 27, e79091. https://doi.org/10.2196/79091.

33. Mavrych V, Yousef EM, Yaqinuddin A, Bolgova O. Large language models in medical education: A comparative cross-platform evaluation in answering histological questions. *Med Educ Online.* **2025**, 30(1), 2534065. https://doi.org/10.1080/10872981.2025.2534065.
34. Fortuna A, Prasetya F, Samala AD, Rawas S, Criollo-C S, Kaya D, et al. Artificial intelligence in personalized learning: A global systematic review of current advancements and shaping future opportunities. *Soc Sci Humanit Open.* **2025**, 12, 102114. https://doi.org/10.1016/j.ssaho.2025.102114.
35. Tran M, Balasooriya C, Jonnagaddala J, Leung GKK, Mahboobani N, Ramani S, et al. Situating governance and regulatory concerns for generative artificial intelligence and large language models in medical education. *NPJ Digit Med.* **2025**, 8, 315. https://doi.org/10.1038/s41746-025-01721-z