Revista Española de Educación Médica

UNIVERSIDAD DE MURCIA

Retroalimentación automatizada mediante inteligencia artificial en simulación clínica: revisión crítica integrativa multifuente de validez, aprendizaje y transferencia.

Artificial intelligence-generated automated feedback in clinical simulation: a multisource integrative critical review of validity, learning, and transfer.

Ana Mikal Toala Mora[1]*, Adriana Denisse García Coello[2]

[1] Carrera de Medicina, Universidad San Gregorio de Portoviejo, Portoviejo, Ecuador; amtoala@sangregorio.edu.ec, https://orcid.org/0000-0002-4301-4264; [2] Adriana Denisse García Coello: adgarcia@sangregorio.edu.ec, https://orcid.org/0000-0002-7044-1894.

* Correspondencia: amtoala@sangregorio.edu.ec

Recibido: 3/9/26; Aceptado: 28/9/26; Publicado: 29/6/26

Resumen.

Objetivo: Analizar críticamente la evidencia empírica sobre retroalimentación automatizada mediante inteligencia artificial (IA) en simulación clínica para estudiantes de medicina, con énfasis en validez, resultados educativos y transferencia. **Metodología**: Se realizó una revisión crítica integrativa multifuente, no sistemática, orientada por el enfoque metodológico de Whittemore y Knafl y por recomendaciones para revisiones integrativas. Se efectuaron búsquedas estructuradas en PubMed/MEDLINE, ERIC y OpenAlex, complementadas con rastreo de citas y contraste con revisiones recientes. Los estudios primarios fueron sometidos a una valoración crítica explícita por dominios de diseño, comparador, independencia del desenlace, reproducibilidad y seguimiento. **Resultados**: Se analizaron 17 estudios empíricos publicados entre 2019 y 2026. La evidencia mostró concordancia elevada con evaluadores humanos en algunas tareas estructuradas, pero también errores de especificidad, sesgo hacia puntuaciones altas, variabilidad entre ejecuciones y limitaciones en conductas comunicativas complejas. Los beneficios educativos se concentraron en desenlaces inmediatos; la transferencia a evaluaciones independientes se documentó solo en un subconjunto. **Conclusiones**: La retroalimentación automatizada muestra utilidad principalmente formativa y para práctica repetida. La evidencia disponible no demuestra equivalencia general con evaluación humana ni permite sostener su uso autónomo en decisiones sumativas de alto impacto. Se requieren validaciones por dominio, desenlaces independientes y seguimiento longitudinal.

Palabras clave: inteligencia artificial, retroalimentación automatizada, simulación clínica, pacientes virtuales, educación médica, evaluación formativa

Abstract.

Objective: To critically analyze empirical evidence on artificial intelligence (AI)-generated automated feedback in clinical simulation for medical students, focusing on validity, educational outcomes, and transfer. **Methods:** A multisource integrative critical review, rather than a systematic review, was conducted following the methodological principles of Whittemore and Knafl and guidance for integrative reviews. Structured searches were performed in PubMed/MEDLINE, ERIC, and OpenAlex, supplemented by citation tracking and comparison with recent reviews. Primary studies underwent an explicit critical appraisal across design, comparator, outcome independence, reproducibility, and follow-up domains. **Results:** Seventeen empirical studies published between 2019 and 2026 were analyzed. High agreement with human

raters was reported for some structured tasks, but specificity errors, overly generous scoring, run-to-run variability, and limitations in complex communication behaviors were also observed. Educational benefits were concentrated in immediate outcomes, while transfer to independent assessments was documented only in a subset. **Conclusions**: Automated feedback currently shows its clearest utility for formative and repeated practice. Available evidence does not establish general equivalence with human assessment or justify autonomous use in high-stakes summative decisions. Domain-specific validation, independent outcomes, and longitudinal follow-up are needed.

Keywords: artificial intelligence, automated feedback, clinical simulation, virtual patients, medical education, formative assessment

1. Introducción

La simulación clínica permite entrenar competencias clínicas en entornos seguros y controlados, con posibilidad de repetición, estandarización y retroalimentación antes de la exposición a situaciones asistenciales de mayor complejidad. En educación médica, los pacientes estandarizados, los simuladores físicos, los pacientes virtuales y los entornos inmersivos no son modalidades equivalentes, sino recursos con diferentes grados de realismo, disponibilidad y capacidad para entrenar dominios específicos. Fierro Manrique et al. destacaron recientemente la complementariedad entre la simulación de alta fidelidad con pacientes estandarizados y los entornos virtuales, señalando ventajas diferenciales en comunicación, realismo emocional, accesibilidad y repetición (1).

La expansión de la inteligencia artificial (IA), y en particular de los modelos de lenguaje de gran escala (LLM), ha añadido una nueva función a la simulación: generar respuestas clínicas dinámicas, interpretar el diálogo del estudiante y producir retroalimentación automática durante o después del encuentro. Una revisión de alcance publicada en esta misma revista mapeó aplicaciones de IA en formación clínica práctica y encontró predominio de escenarios simulados y resultados inmediatos (2). Revisiones internacionales recientes también describen crecimiento rápido de pacientes virtuales generativos y aplicaciones de IA en habilidades clínicas, pero coinciden en la heterogeneidad metodológica y en la limitada evidencia de transferencia a contextos clínicos reales (3-5). Trabajos recientes han descrito la expansión de la IA generativa, el uso pedagógico de ChatGPT, la generación de casos con LLM y nuevas competencias docentes vinculadas a IA; en conjunto, muestran un campo de rápida evolución y resultados todavía predominantemente tempranos (51-54).

En ese panorama, la retroalimentación automatizada merece un análisis específico. El retroalimentación no es simplemente información posterior al desempeño: su valor depende de que permita reconocer una brecha, comprenderla y actuar sobre ella. Los marcos clásicos de retroalimentación destacan la especificidad, la comparación con un estándar y la orientación hacia acciones de mejora (6, 7). En simulación, esta función se vincula con la práctica deliberada, que requiere objetivos definidos, repetición y retroalimentación correctiva (8). Desde la teoría de carga cognitiva, además, la cantidad y complejidad de la retroalimentación deben ajustarse al nivel del aprendiz para evitar sobrecarga extrínseca (9, 10). Por tanto, la fluidez lingüística de un modelo no demuestra por sí misma que su juicio sea válido, accionable o pedagógicamente útil.

La revisión de alcance de Bonilla Mejia et al. (2) incluyó la evaluación y la retroalimentación automatizadas como uno de varios dominios de aplicación de la IA, mientras otras revisiones recientes se han centrado en retroalimentación generada por LLM o en pacientes virtuales de forma amplia (14-17, 44). Permanece menos resuelta una pregunta específica: si la retroalimentación automatizada produce cambios que se mantienen cuando el estudiante es evaluado fuera de la propia plataforma de IA. Por ello, el aporte diferencial de esta revisión es separar tres niveles que a

menudo se mezclan: concordancia con evaluadores humanos, calidad pedagógica de la retroalimentación y transferencia hacia evaluaciones independientes como OSCE o pacientes estandarizados. Este foco complementa, y no duplica, los mapeos generales disponibles.

El objetivo de este artículo es analizar críticamente la evidencia empírica sobre retroalimentación y evaluación automatizadas mediante IA dentro de simulaciones clínicas dirigidas a estudiantes de medicina. Se plantean cuatro preguntas: (1) ¿qué evidencias existen sobre exactitud, concordancia y estabilidad del retroalimentación automatizada?; (2) ¿qué resultados educativos se han asociado con su uso?; (3) ¿en qué medida los beneficios se transfieren a evaluaciones independientes de la plataforma de IA?; y (4) ¿qué salvaguardas pedagógicas y operativas se desprenden de la evidencia disponible?

El análisis se estructuró desde cuatro marcos teóricos definidos a priori. Primero, la teoría de carga cognitiva orientó el análisis de cantidad, secuencia y complejidad de la retroalimentación (9,10). Segundo, los criterios de buena evaluación y el enfoque argumentativo de validez permitieron distinguir entre exactitud de una salida y suficiencia de evidencia para el uso evaluativo pretendido (11,12). Tercero, la evaluación programática aportó el principio de que decisiones de alto impacto requieren múltiples puntos de información y juicio experto, mientras un dato aislado puede conservar valor formativo (13). Cuarto, la práctica deliberada sustentó el análisis de repetición, retroalimentación correctiva y progresión del desempeño (8). Estos marcos justificaron los dominios examinados: calidad y accionabilidad de la retroalimentación, validez de la evaluación automatizada, condiciones de práctica, carga cognitiva y transferencia a evaluaciones independientes. La necesidad de un foco específico también se ha reforzado por síntesis recientes. Revisiones sobre pacientes virtuales impulsados por IA para comunicación, entrevista y razonamiento clínico han identificado beneficios potenciales, pero también falta de marcos teóricos consistentes, medidas estandarizadas y evidencia longitudinal (14-17). La presente revisión no pretende reemplazar esas síntesis amplias, sino examinar críticamente el componente de retroalimentación/evaluación y contrastar los estudios primarios localizados con dichas revisiones.

2. Métodos

2.1. Diseño y enfoque de la revisión

Se realizó una revisión crítica integrativa con búsqueda estructurada multifuente. El trabajo se orientó por el enfoque de Whittemore y Knafl, que organiza la revisión integrativa en definición del problema, búsqueda de literatura, evaluación de los datos, análisis y presentación de la síntesis, y por recomendaciones metodológicas contemporáneas para revisiones de literatura (48,49). Este diseño se eligió porque permite integrar estudios experimentales, cuasiexperimentales, de validación y métodos mixtos alrededor de un mecanismo educativo específico -la retroalimentación automatizada dentro de la simulación clínica-. El manuscrito no se presenta como revisión sistemática ni de alcance; por tanto, no reclama exhaustividad ni cumplimiento de PRISMA-ScR. Las inferencias se restringen a los estudios primarios identificados mediante la estrategia descrita y a su contextualización con síntesis secundarias recientes.

2.2. Pregunta de revisión y elegibilidad

La pregunta fue: ¿qué evidencia empírica existe sobre validez, calidad, resultados educativos y transferencia de la retroalimentación o evaluación automatizadas generadas por IA dentro de experiencias de simulación clínica destinadas a estudiantes de medicina? Se incluyeron publicaciones empíricas entre enero de 2019 y el 3 de septiembre de 2026 que: (a) estuvieran dirigidas a estudiantes de medicina o a un contexto explícito de educación médica; (b) emplearan un paciente virtual, paciente simulado digital, interacción con paciente estandarizado registrada, agente conversacional o escenario clínico interactivo; y (c) utilizaran IA para generar

retroalimentación, calificar el desempeño o comparar una evaluación automatizada con un referente humano. Se admitieron estudios de validación tecnológica cuando evaluaban explícitamente la calidad de la retroalimentación o puntuación destinada al aprendiz. Se excluyeron editoriales, cartas, protocolos, revisiones secundarias como estudios primarios, simulaciones sin retroalimentación/evaluación automatizada y aplicaciones de IA usadas solo para generar contenido sin interacción clínica.

2.3. Búsqueda bibliográfica

La identificación se realizó mediante búsquedas estructuradas y complementarias en PubMed/MEDLINE, ERIC y OpenAlex. En cada fuente se combinaron conceptos de población/contexto (medical students, medical education), simulación (virtual patient, simulated patient, standardized patient, clinical simulation), IA (artificial intelligence, generative AI, large language model, GPT) y retroalimentación/evaluación (feedback, automated assessment, scoring, evaluation). No se aplicó restricción por disponibilidad de texto completo. Se consideraron publicaciones en inglés o español; los trabajos en otros idiomas fueron elegibles si disponían de información suficiente en inglés para valorar los criterios. Se efectuó rastreo retrospectivo de referencias y prospectivo de citaciones. Como auditoría de cobertura, los estudios recuperados se contrastaron con cuatro síntesis recientes (14-17) que incluyeron bases adicionales. Esta triangulación se utilizó para detectar omisiones potenciales y no como sustituto de una búsqueda sistemática directa en bases de suscripción. La búsqueda original se actualizó el 3 de septiembre de 2026; las cadenas y el procedimiento de recuperación se presentan en el Anexo 1.

Tabla 1. Fuentes de información y estrategia de identificación de la revisión crítica integrativa.

Fuente	Función en la revisión	Bloques de búsqueda / estrategia
PubMed/ MEDLINE	Identificación biomédica principal	medical education/student* + virtual/simulated/standardized patient + artificial intelligence/LLM/GPT + retroalimentación/assessment/scoring
ERIC	Cobertura educativa complementaria	medical education/student* + simulation/virtual patient + artificial intelligence/LLM + retroalimentación/assessment
OpenAlex	Cobertura multidisciplinar y detección de trabajos no recuperados en PubMed/ERIC	Combinaciones por título/tema de AI-generated retroalimentación, virtual patient, standardized patient, clinical simulation y medical education
Rastreo de citas	Búsqueda retrospectiva y prospectiva	Referencias de estudios incluidos, artículos relacionados y registros de citación
Síntesis multibase recientes (14-17)	Auditoría de cobertura y contextualización	Contraste con revisiones que emplearon Scopus, Web of Science, CINAHL, PsycINFO, ERIC y otras bases

2.4. Proceso de selección y control de duplicados

Los registros se depuraron por coincidencia de DOI y título. En una primera fase se revisaron título y resumen según los criterios de elegibilidad; los trabajos potencialmente pertinentes pasaron a evaluación a texto completo. Los artículos identificados por rastreo de citas y OpenAlex se sometieron a los mismos criterios. Las revisiones secundarias no se contabilizaron entre los estudios primarios y se utilizaron únicamente para contraste de cobertura y discusión. La autora principal

(AMTM) realizó el cribado inicial y la extracción de datos mediante una hoja predefinida. La segunda autora (ADGC) verificó de manera independiente la elegibilidad de los estudios incluidos y las variables críticas extraídas -diseño, muestra, intervención, comparador, desenlace principal y DOI-. Las discrepancias se resolvieron por consenso. Este procedimiento no equivale a cribado independiente por duplicado desde la identificación y se reconoce como limitación.

La revisión se desarrolló inicialmente como revisión crítica y el registro histórico de búsqueda no conservó un recuento exportable por fuente de todos los registros recuperados y excluidos. Para evitar reconstruir cifras retrospectivas no verificables, no se presentan números que no puedan auditarse. Sí se conserva la trazabilidad de los 17 estudios primarios incluidos, sus DOI, diseño, muestra, desenlaces y motivo de inclusión (Tabla 2). Esta limitación de trazabilidad cuantitativa se incorpora expresamente en la sección de limitaciones y constituye una diferencia respecto de una revisión sistemática.

2.5. Extracción, verificación y síntesis de la evidencia

Para cada estudio se extrajeron año, país, diseño, participantes, modalidad de simulación, tipo de IA, función del sistema, naturaleza de la retroalimentación o evaluación, comparador, desenlaces y hallazgos relevantes. Los dominios de síntesis se definieron antes de interpretar los resultados y se fundamentaron en los marcos expuestos en la introducción: (1) validez, concordancia, reproducibilidad y uso previsto; (2) especificidad, accionabilidad y calidad pedagógica de la retroalimentación; (3) resultados educativos y condiciones de práctica deliberada; (4) carga cognitiva y experiencia del estudiante; y (5) transferencia a una evaluación independiente de la propia plataforma. La heterogeneidad de diseños, competencias, escalas y comparadores impidió una síntesis cuantitativa agregada, por lo que se realizó análisis narrativo por dominios.

2.6. Valoración crítica de la evidencia

Dada la heterogeneidad de diseños, se realizó una valoración crítica por dominios, inspirada en los principios de evaluación de estudios heterogéneos del Mixed Methods Appraisal Tool (MMAT) (50), sin convertir sus criterios en una puntuación numérica global. Para cada estudio se consideraron cinco dominios: (1) adecuación del diseño y presencia de comparador o referente; (2) suficiencia y representatividad de la muestra para el objetivo; (3) independencia del desenlace respecto de la propia plataforma de IA; (4) reproducibilidad o repetición de la evaluación; y (5) duración del seguimiento. La lectura global se clasificó como mayor, intermedia o exploratoria para orientar la interpretación, no como nivel formal de certeza. Los estudios con comparador, desenlace independiente y diseño experimental recibieron mayor peso interpretativo; los pilotos pequeños, estudios perceptuales y series sin control se consideraron exploratorios. La Tabla 3 presenta esta valoración.

2.7. Transparencia y reproducibilidad

Para favorecer la reproducibilidad se documentaron las fuentes consultadas, las cadenas de búsqueda, los criterios de elegibilidad, las variables de extracción y los dominios de valoración crítica. La figura 1 conserva una función descriptiva del proceso y no se presenta como diagrama PRISMA. La trazabilidad cuantitativa retrospectiva por fuente es incompleta, por lo que no se reconstruyeron cifras no verificables. Esta decisión prioriza transparencia sobre una apariencia de exhaustividad que el diseño original no puede sostener.

2.8. Consideraciones éticas

El estudio utilizó exclusivamente información procedente de publicaciones disponibles y no involucró participantes humanos, historias clínicas identificables ni datos personales; por ello no

requirió aprobación de un comité de ética. La revisión se orientó por principios de transparencia, trazabilidad de fuentes y verificación de los datos bibliográficos.

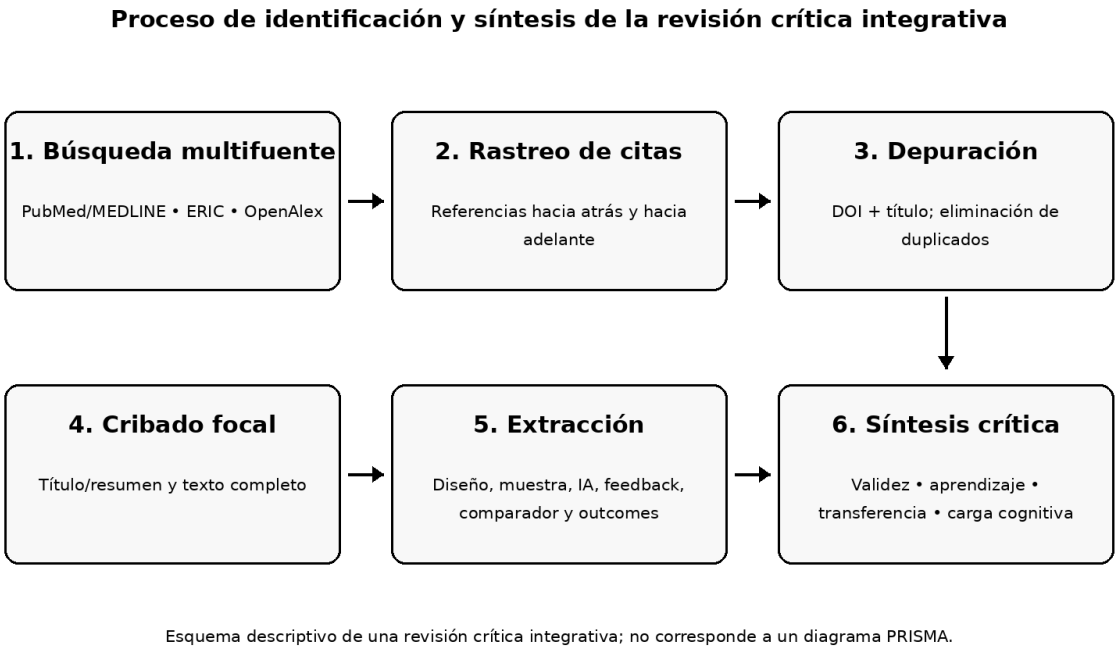

Figura 1. Proceso de identificación y síntesis de la revisión crítica integrativa.

3. Resultados

3.1. Características generales

La síntesis primaria incluyó 17 estudios empíricos que cumplieron los criterios focales. Uno fue publicado en 2019 y los restantes entre 2024 y 2026. Además de los estudios previamente localizados, la búsqueda multifuente incorporó un ensayo aleatorizado sobre simulación con LLM y retroalimentación estructurado, una evaluación multimetric de ChatGPT como paciente estandarizado y evaluador, un piloto de escucha ambiental para retroalimentación en encuentros simulados y un estudio de evaluación de transcripciones de encuentros con pacientes estandarizados mediante varios LLM (18-21). Salvo el sistema previo a la actual generación de LLM descrito por Maicher et al. (22), los trabajos recientes utilizaron predominantemente LLM o IA generativa. La anamnesis, comunicación y razonamiento clínico fueron los dominios más frecuentes. La Tabla 2 resume los estudios primarios incluidos.

3.2. Valoración crítica de la evidencia

La valoración por dominios mostró que la evidencia de mayor peso interpretativo procede de los ensayos aleatorizados y de estudios con comparadores o desenlaces independientes, mientras varios pilotos, estudios perceptuales y series pre-post aportan evidencia exploratoria. Ningún estudio aportó seguimiento prolongado con pacientes reales. Por ello, los resultados siguientes se presentan jerarquizados por diseño y por independencia del desenlace, evitando otorgar el mismo peso a todos los estudios (tabla 3).

Tabla 2. Estudios empíricos incluidos sobre retroalimentación o evaluación automatizados mediante IA en simulación clínica.

Estudio	País	Diseño/muestra	Modalidad	Retroalimentación/ evaluación	Hallazgo principal	Lectura crítica
Maicher et al., 2019 (22)	EE. UU.	Validación; 102 encuentros (20 analizados)	Paciente estandarizado virtual 3D con IA	Clasificación automática de preguntas y retroalimentación inmediato	Exactitud global 87% frente a 90% humana; alta fiabilidad en 19/21 categorías.	Demuestra factibilidad de evaluación automatizada previa a LLM.
Holderried et al., 2024 (23)	Alemania	Prospectivo; 106 conversaciones, 1894 pares pregunta-respuesta	Paciente simulado con GPT-4	Feedback estructurado sobre anamnesis	>99% de respuestas médicamente plausibles; κ=0,832 con evaluador humano; κ<0,6 en 8/45 categorías.	Buena concordancia global, pero especificidad desigual por categoría.
Yamamoto et al., 2024 (24)	Japón	No aleatorizado; 35 intervención/110 control histórico	Paciente simulado con LLM	Feedback posterior a entrevistas	OSCE preclínico: 28,1 vs 27,1; p=0,01. Limitación en comunicación no verbal.	Evidencia de transferencia a evaluación estandarizada, con diseño no aleatorizado.
Scherr et al., 2025 (25)	EE. UU.	Evaluación experimental; 360 simulaciones	ChatGPT-4	Feedback clínico dentro de simulación de manejo	100% exactitud en parámetros básicos; retroalimentación integral en 78%; retroalimentación diferido en 55%.	La adherencia a reglas pedagógicas avanzadas no fue consistente.
Weisman et al., 2025 (26)	EE. UU.	Estudio multifásico de desarrollo	Paciente virtual GPT-4 para mamografía anormal	Feedback escrito sobre comunicación/SPIKES	Identificó fortalezas y áreas de mejora, pero ocasionalmente clasificó erróneamente la adherencia al protocolo.	El diseño de prompts y la rúbrica condicionan la calidad del retroalimentación.
Laverde et al., 2025 (27)	Colombia/ Luxemburgo	Desarrollo y evaluación de chatbot	Agentes LLM para anamnesis	Feedback en tiempo real y práctica repetible	Alta usabilidad reportada; respuestas plausibles y práctica asincrónica flexible.	Fortaleza en escalabilidad; requiere validación externa del desempeño.
Suárez-García et al., 2025 (28)	México	Pre-post; 30 estudiantes	GPT-4o, 10 escenarios virtuales	Feedback automático en lenguaje natural	Desempeño aumentó 36,7 puntos; 0% a 70% de alto desempeño; postest con pacientes estandarizados humanos.	Señal fuerte de transferencia, pero sin grupo control.

Jacobs et al., 2025 (29)	Reino Unido	Métodos mixtos; 18 participantes (15 estudiantes)	Paciente virtual con LLM y voz	Feedback inmediato	Alta motivación, usabilidad y seguridad psicológica; fidelidad moderada.	El retroalimentación fue valorado, pero el realismo no igualó al paciente humano.
Luo et al., 2025 (30)	China	ECA; n=84	Paciente digital con LLM y voz	Feedback adaptativo	Mejora de 10,50 puntos en anamnesis frente a método tradicional; mayor empatía.	Uno de los diseños más robustos; contexto especializado en oftalmología.
Byun et al., 2026 (31)	Corea	Estudio de percepciones; n=21	GPT-4, 7 pacientes virtuales + evaluador virtual	Diálogo reflexivo inmediato + informe escrito	Mayor valoración de especificidad/personalización y autoeficacia frente a métodos convencionales.	Resultados predominantemente perceptuales; requiere validación objetiva.
Borg et al., 2026 (32)	Suecia	Cuasiexperimental; n=115	Paciente virtual robótico + IA	Feedback postconsulta automatizado	OSCE: 7,39 vs 6,68; diferencia 0,70; d=0,74; aprobados 96,7% vs 79,6%.	Asociación favorable con desempeño independiente; no todas las subescalas mejoraron.
Sindhvananda et al., 2026 (33)	Tailandia	Métodos mixtos pre-post; n=25	Paciente simulado con IA para cefalea	Feedback automático basado en rúbrica	OSCE +22,5 puntos (d=1,93); CRI-HT +8,8 (d=1,81); diálogo percibido como algo robótico.	Efectos grandes, pero muestra pequeña y ausencia de control paralelo.
Chan et al., 2026 (34)	Australia	Piloto prospectivo; 43 voluntarios + 8 remediación	Paciente virtual con IA, 9 urgencias	Feedback inmediato	Sin mejora clara en voluntarios; subgrupo de remediación mejoró OSCE 37 a 67 (p=0,03).	Resultados exploratorios; posible mayor utilidad en necesidades específicas.
Brügge et al., 2024 (18)	Alemania	ECA doble ciego; 21 estudiantes	Paciente LLM; 4 historias clínicas	Feedback estructurado generado por LLM	Grupo retroalimentación: 3,60 vs 3,02 en CRI-HTI al final; $F(1,18)=4,44$; $p=0,049$; $\eta^2p=0,198$.	Evidencia experimental directa de efecto adicional del retroalimentación, pero muestra pequeña.
Wang et al., 2025 (19)	China	Evaluación multimetric; múltiples ejecuciones por	ChatGPT como paciente estandarizado y	Puntuación de la consulta y retroalimentación según criterios	Diferenció consultas pobres de moderadas/buenas; prompt engineering redujo desviación y	Muestra sensibilidad a prompt y tendencia inicial a sobrepuntuar;

		escenario	evaluador		variabilidad de puntuación.	útil para estudiar validez del scoring.
Snedegar et al., 2026 (20)	EE. UU.	Piloto; 10 evaluaciones de estudiantes	Encuentro con paciente simulado + escucha ambiental IA	Feedback postencuentro y metas SMART	Alta aceptación y alineación percibida con conductas observables; se solicitó mayor contextualización y ejemplos.	Factibilidad prometedora; n muy pequeño y outcomes principalmente perceptivos.
Nolan y Burke, 2026 (21)	EE. UU.	Validación; 12 encuentros estandarizados	Transcripciones de encuentros SP evaluadas por ChatGPT-4o y Gemini	Scoring + retroalimentación narrativo	Anamnesis: κ=0,84-0,89; ICC=0,93-0,97. Presentación: κ=0,62-0,72; ICC=0,61-0,83.	Concordancia alta en elementos estructurados y menor en componentes complejos; muestra pequeña.

Tabla 3. Valoración crítica por dominios de los 17 estudios primarios incluidos. "Peso interpretativo" orienta la síntesis narrativa y no constituye una puntuación formal de certeza.

Estudio	Diseño/ comparador	Desenlace independiente	Reproducibilidad/seguimiento	Peso interpretativo
Maicher 2019 (22)	Validación con referente humano	No	Repetición por categorías; sin seguimiento	Intermedio
Holderried 2024 (23)	Prospectivo con evaluador humano	No	Múltiples conversaciones; sin seguimiento	Intermedio
Yamamoto 2024 (24)	No aleatorizado; control histórico	Sí, OSCE	Seguimiento corto	Intermedio
Scherr 2025 (25)	Evaluación experimental	No	360 simulaciones; sin seguimiento	Intermedio
Weisman 2025 (26)	Desarrollo multifásico	No	Ejecuciones repetidas; sin seguimiento	Exploratorio-intermedio
Laverde 2025 (27)	Desarrollo/ evaluación sin control	No	Validación externa limitada	Exploratorio
Suárez-García 2025 (28)	Pre-post sin control	Sí, paciente estandarizado	Seguimiento inmediato	Intermedio
Jacobs 2025 (29)	Métodos mixtos; muestra pequeña	No	Predominio perceptual	Exploratorio
Luo 2025 (30)	ECA con comparador	Parcial	Seguimiento corto	Mayor
Byun 2026 (31)	Percepciones; n pequeño	No	Sin seguimiento	Exploratorio
Borg 2026 (32)	Cuasiexperimental	Sí, OSCE	Seguimiento inmediato	Intermedio-mayor
Sindhvananda 2026 (33)	Pre-post; métodos mixtos	Sí, OSCE	Muestra pequeña; sin control	Intermedio
Chan 2026 (34)	Piloto prospectivo	Sí, OSCE	Subgrupo pequeño	Exploratorio-intermedio
Brügge 2024 (18)	ECA doble ciego; n pequeño	No/medida estructurada	Seguimiento corto	Mayor
Wang 2025 (19)	Validación multiejecución	Referente de calidad	Sí; sensibilidad a prompt	Intermedio
Snedegar 2026 (20)	Piloto; n=10	No	Outcomes perceptivos	Exploratorio
Nolan y Burke 2026 (21)	Validación con evaluadores	No	Dos LLM; n=12 encuentros	Intermedio

3.3. Validez, concordancia y calidad de la retroalimentación automatizada

Los estudios de validación muestran que la calidad global puede ser alta para tareas estructuradas, pero no uniforme. Wang et al. (19) mostraron que ChatGPT podía diferenciar desempeños pobres de moderados/buenos y que la ingeniería de prompts reducía de forma

marcada la desviación de las puntuaciones, aunque el modelo tendía inicialmente a sobrepuntuar consultas de baja calidad. Nolan y Burke (21), al evaluar transcripciones de 12 encuentros con pacientes estandarizados, observaron acuerdo alto para elementos de anamnesis (κ=0,84-0,89; ICC=0,93-0,97) y menor para componentes de presentación clínica (κ=0,62-0,72; ICC=0,61-0,83). Entre los estudios previos, Maicher et al. (22) comunicaron una exactitud automatizada del 87%, próxima a la humana (90%), mientras Holderried et al. (23) hallaron respuestas médicamente plausibles en más del 99% y concordancia casi perfecta global con un evaluador humano (κ=0,832), aunque ocho de 45 categorías presentaron κ inferior a 0,6. En conjunto, la concordancia global no excluye fallos relevantes por dominio.

Los problemas no se limitan a la exactitud factual. Scherr et al. (25), al ejecutar 360 simulaciones de manejo clínico, encontraron exactitud médica en parámetros básicos, pero solo 78% de las simulaciones proporcionaron retroalimentación integral y 55% respetaron la condición de retroalimentación diferida. Weisman et al. (26) describieron errores al juzgar la adhesión al protocolo SPIKES y variabilidad cuando el mismo texto era evaluado en ejecuciones repetidas. Wang et al. (19) también identificaron puntuaciones inicialmente demasiado altas y respuestas que incorporaban información no solicitada. Estos hallazgos sugieren que la validez de la retroalimentación debe analizarse por dominio, rúbrica, versión del modelo y uso previsto, y no inferirse de la fluidez lingüística.

Dentro de los estudios primarios incluidos, se observaron patrones de sobrepuntuación, omisión de debilidades y variabilidad entre ejecuciones, especialmente en Wang et al. (19), Weisman et al. (26) y Nolan y Burke (21). Estos hallazgos no equivalen a una demostración de invalidez general; indican que la concordancia debe interpretarse por dominio y que una salida lingüísticamente plausible no constituye por sí sola evidencia suficiente para un uso evaluativo determinado.

3.4. Resultados educativos y experiencia del estudiante

La mayoría de los estudios que incorporaron estudiantes informó una experiencia favorable en términos de disponibilidad, seguridad psicológica, personalización o utilidad percibida. Snedegar et al. (20), en un piloto con encuentros simulados registrados mediante escucha ambiental, encontraron que los participantes percibían la retroalimentación como vinculado a conductas observables y útil para generar metas de mejora, aunque solicitaron mayor contextualización y ejemplos más específicos. Jacobs et al. (29) observaron alta motivación, usabilidad y seguridad psicológica, pese a fidelidad moderada del paciente virtual, y Byun et al. (31) encontraron que los estudiantes valoraron la retroalimentación del evaluador virtual como específico y personalizado. Estas percepciones son relevantes para la aceptabilidad, pero no sustituyen medidas objetivas de aprendizaje.

Los estudios con medidas objetivas muestran señales favorables, aunque con diseños heterogéneos. Brügge et al. (18), en un ensayo aleatorizado de 21 estudiantes, encontraron que el grupo que recibió retroalimentación generada por LLM tras cuatro simulaciones mejoró más el desempeño de toma de decisiones clínicas que el grupo con simulación sin retroalimentación (η^2 parcial=0,198), con beneficios en creación de contexto y aseguramiento de información. Yamamoto et al. (24) reportaron una diferencia significativa en entrevista del OSCE preclínico a favor del grupo entrenado con pacientes simulados por IA. Luo et al. (30) encontraron en un ensayo aleatorizado una mejora de 10,50 puntos en historia clínica oftalmológica frente a métodos tradicionales. Sindhvananda et al. (33) observaron incrementos pre-post en OSCE y razonamiento clínico después de practicar cefaleas con retroalimentación basado en rúbrica. En contraste, Chan et al. (34) no identificaron una mejoría clara en el grupo voluntario, aunque el pequeño subgrupo de remediación mostró un cambio favorable.

Borg et al. (32) y Yamamoto et al. (24) evaluaron desempeño mediante estaciones estandarizadas externas al agente. Sindhvananda et al. (33) utilizaron OSCE y razonamiento clínico, y Chan et al. (34) exploraron desempeño tipo OSCE tras entrenamiento con paciente virtual. Sin embargo, ningún estudio localizado demostró transferencia sostenida a resultados clínicos con pacientes reales ni retención a largo plazo. Esta brecha delimita la fuerza de cualquier afirmación sobre competencia clínica más allá del entorno educativo inmediato.

3.5. Transferencia a evaluaciones independientes

El criterio más exigente es si la mejora producida por retroalimentación de IA aparece posteriormente en una evaluación independiente de la propia herramienta. Al menos cinco estudios utilizaron OSCE, pacientes estandarizados humanos u otra evaluación externa como outcome (24, 28, 32-34). Suárez-García et al. (28), por ejemplo, utilizaron pacientes estandarizados humanos en el postest después de entrenamiento virtual, mientras Borg et al. (32) evaluaron mediante un OSCE independiente con paciente estandarizado y evaluador cegado. Estas señales de transferencia son más informativas que las puntuaciones generadas por el mismo sistema que entrena al estudiante, aunque continúan concentradas en resultados inmediatos o de corto plazo.

Borg et al. (32) y Yamamoto et al. (24) también evaluaron desempeño mediante estaciones estandarizadas externas al agente. Sindhvananda et al. (33) utilizaron OSCE pre-post, mientras Chan et al. (34) exploraron desempeño en un programa de preparación para OSCE. Estos estudios acercan la evidencia a una transferencia educativa, pero el horizonte temporal sigue siendo corto. No se identificaron estudios que demostraran retención a varios meses, transferencia consistente al encuentro con pacientes reales o efectos sobre seguridad/calidad asistencial.

Tabla 4. Síntesis de la fuerza y las brechas de evidencia por dominio.

Dominio	Señal de evidencia	Limitación recurrente	Interpretación para la práctica
Exactitud/concordancia	Concordancia alta en tareas estructuradas y sensibilidad a ingeniería de prompts (19,21-23).	Variación por dominio, sobrepuntuación, especificidad desigual y variabilidad entre ejecuciones (19,21,23,26,35).	Validar por rúbrica y por dominio antes de usar el sistema.
Experiencia del estudiante	Aceptabilidad y percepción de utilidad en varios estudios, incluido retroalimentación tras encuentros simulados registrados (20,29,31).	Fidelidad moderada; retroalimentación percibido no equivale a retroalimentación válido.	Útil para práctica formativa, no suficiente para inferir competencia.
Desempeño educativo	Mejoras en razonamiento, anamnesis, comunicación u OSCE en ECA, estudios controlados y pre-post (18,24,28,30,32-33).	Diseños, comparadores y escalas heterogéneos; tamaños muestrales pequeños en varios trabajos.	Prometedor como complemento; no permite estimar un efecto agregado.

Transferencia	OSCE o pacientes estandarizados independientes en un subconjunto (24,28,32-34).	Seguimiento corto; sin resultados con pacientes reales.	Priorizar evaluaciones externas y longitudinales en próximos estudios.
Uso sumativo	No se identificó evidencia robusta para decisiones de alto impacto.	Problemas de estabilidad, omisiones y juicio contextual.	Mantener supervisión humana y evitar certificación autónoma.

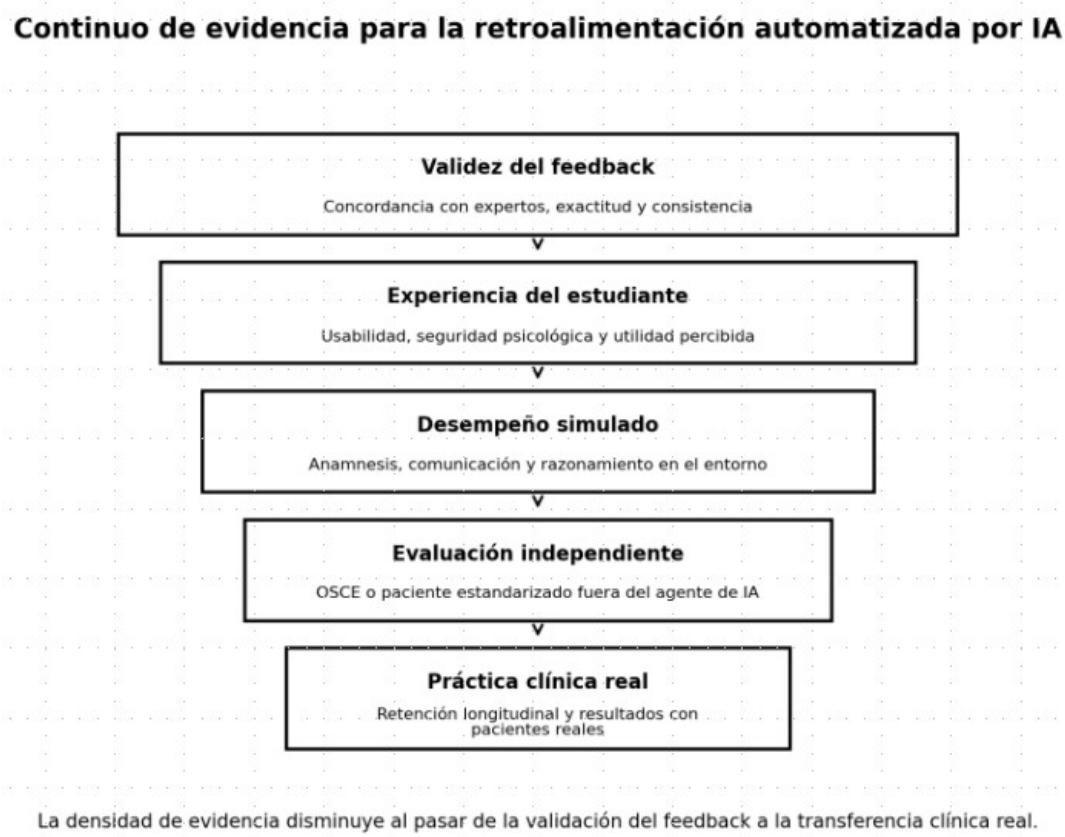

Figura 2. Continuo de madurez de la evidencia sobre retroalimentación automatizada mediante IA. Nota: el esquema representa una síntesis conceptual de los hallazgos. No implica que los dominios sean mutuamente excluyentes.

3.5. Triangulación con síntesis multibase recientes

La triangulación con revisiones publicadas entre 2024 y septiembre de 2026 mostró concordancia general con los patrones observados en los 17 estudios primarios: predominio de anamnesis y comunicación, resultados tempranos favorables, alta heterogeneidad, escasa estandarización de outcomes y evidencia limitada de transferencia longitudinal (14-17). La incorporación de búsquedas en ERIC y OpenAlex permitió localizar estudios adicionales no presentes en la síntesis focal previa, lo que refuerza la utilidad de una estrategia multifuente para un campo que se publica tanto en revistas biomédicas como educativas y tecnológicas.

3.6. Literatura de contextualización

Las revisiones secundarias y otros trabajos no incluidos entre los 17 estudios primarios se utilizaron únicamente para contextualizar los hallazgos y no para estimar los resultados de la síntesis primaria. Estas fuentes informan el marco pedagógico, la evolución del campo y los problemas de ética, gobernanza, implementación y obsolescencia tecnológica (14-17,35-47,51-54).

4. Discusión

4.1. Hallazgos principales

Los estudios localizados muestran una tensión central: la retroalimentación automatizada por IA puede ampliar las oportunidades formativas y, en ciertos diseños, asociarse con mejoras objetivas; sin embargo, la evidencia de validez y eficacia es dependiente de la tarea. Brügge et al. (18) y Borg et al. (32) proporcionan señales experimentales o cuasiexperimentales de beneficio, mientras Wang et al. (19), Holderried et al. (23), Weisman et al. (26) y Nolan y Burke (21) muestran que el rendimiento evaluativo varía por dominio, prompt, rúbrica y complejidad de la conducta. Por ello, el hallazgo más consistente no es la equivalencia general con expertos, sino una utilidad potencialmente alta en tareas estructuradas y delimitadas. La literatura publicada después de la búsqueda de Bonilla Mejia et al. (2) refuerza la importancia de la retroalimentación como subcampo específico. Los nuevos estudios no solo describen pacientes virtuales, sino que separan progresivamente tres preguntas: si la IA representa adecuadamente al paciente, si evalúa de forma consistente el desempeño y si la retroalimentación modifica un outcome independiente. Brügge et al. (18), Snedegar et al. (20), Borg et al. (32), Sindhvananda et al. (33) y Nolan y Burke (21) ilustran esta transición desde factibilidad hacia validación y resultados educativos, aunque sin resolver todavía la transferencia a práctica clínica real.

4.2. De retroalimentación generativo a retroalimentación pedagógicamente útil

La calidad de la retroalimentación no depende únicamente de su personalización. Los marcos introducidos previamente permiten interpretar por qué algunos sistemas producen resultados útiles en tareas estructuradas y menos consistentes en competencias complejas. La práctica deliberada requiere objetivos claros, ejecución repetida y retroalimentación específica que permita corregir errores (8). La IA puede reducir el tiempo entre desempeño y retroalimentación y facilitar múltiples intentos, pero estas ventajas operativas no equivalen automáticamente a una mejora educativa si la información es imprecisa, difícil de priorizar o no conduce a una nueva oportunidad de práctica. Los hallazgos de Holderried et al. (23), Wang et al. (19) y Weisman et al. (26) ilustran por qué las rúbricas explícitas y el control del prompt pueden actuar como mecanismos de calibración. Cuando el sistema debe decidir si una conducta concreta ocurrió, la concordancia puede ser alta; cuando el juicio exige integrar tono, contexto, empatía o prioridades clínicas, la consistencia disminuye. La calibración técnica, por tanto, debe considerarse una condición para interpretar la retroalimentación y no una garantía de validez educativa completa. La teoría de carga cognitiva aporta una segunda lente interpretativa. El retroalimentación inmediata puede facilitar el aprendizaje, pero una respuesta extensa, no priorizada o contradictoria puede aumentar carga extrínseca. Los marcos de carga cognitiva en educación médica y simulación recomiendan ajustar la complejidad y el apoyo al nivel del aprendiz (9,10). Los estudios incluidos no compararon de forma suficiente distintas dosis o formatos de retroalimentación generativo; por ello, limitar y jerarquizar recomendaciones debe entenderse como una hipótesis pedagógica razonable que necesita evaluación experimental específica.

4.3. La transferencia debe ser el criterio de madurez

Una brecha importante de los estudios localizados es la evidencia limitada de transferencia. Revisiones previas sobre pacientes virtuales han mostrado que los resultados suelen concentrarse

en conocimiento, habilidades inmediatas o percepción (36,37). En la evidencia focal, algunos trabajos avanzan hacia OSCE o pacientes estandarizados independientes (24,28,32-34), pero el seguimiento continúa siendo corto. La ausencia de resultados longitudinales con pacientes reales impide asumir que una mejora durante simulación o evaluación estandarizada se mantenga posteriormente en la práctica clínica.

No obstante, incluso un OSCE inmediato sigue siendo un resultado intermedio. Desde la perspectiva del aprendizaje situado, la competencia clínica emerge de la interacción entre persona, tarea y contexto (38). Por ello, el desempeño en un entorno digital no debe asumirse equivalente al desempeño sostenido con pacientes reales. Estudios futuros podrían fortalecer esta inferencia mediante seguimiento longitudinal, evaluación por observadores independientes y desenlaces clínicos supervisados cuando sea metodológica y éticamente apropiado.

4.4. Implementación, costes y equidad

La escalabilidad es una de las razones más citadas para automatizar retroalimentación. Los agentes conversacionales y sistemas de transcripción pueden multiplicar oportunidades de práctica y disminuir parte de la dependencia de observación docente. Sin embargo, los estudios focales rara vez incorporaron análisis económicos comparativos. La literatura sobre costos de simulación muestra que la sostenibilidad depende de infraestructura, licencias, mantenimiento, desarrollo de casos y tiempo docente, además del precio de la tecnología (39). Por tanto, la evidencia revisada permite plantear la escalabilidad como una hipótesis operativa plausible, no como costo-efectividad demostrada. En instituciones con recursos limitados, los sistemas basados en texto podrían requerir menos infraestructura que experiencias inmersivas complejas, pero esta posibilidad depende de conectividad, licenciamiento, disponibilidad lingüística y soporte local. La equidad debe considerarse como un dominio a evaluar, no como un beneficio asumido. Los estudios localizados aportan poca evidencia comparativa sobre desempeño de la retroalimentación entre grupos lingüísticos, culturales o socioeconómicos.

4.5. Ética, gobernanza y uso en evaluación

El uso de IA para retroalimentación académica introduce cuestiones de privacidad, sesgo, opacidad y responsabilidad. Masters (40) ha propuesto principios de transparencia, supervisión humana, protección de datos y atención a sesgos en educación de profesiones de la salud. En esta revisión, estos principios se emplean como marco interpretativo de riesgo y no como recomendaciones derivadas causalmente de los estudios incluidos. La intensidad de las salvaguardas debería guardar relación con el uso previsto: un sistema para práctica formativa de bajo impacto plantea exigencias distintas a uno utilizado para decisiones de progresión o certificación. La evidencia localizada es más compatible con un uso formativo que con una sustitución autónoma de la evaluación humana. Esta interpretación se apoya en las limitaciones de concordancia observadas y en los principios de validez y evaluación programática: un dato aislado puede ser útil para orientar aprendizaje, mientras decisiones de alto impacto requieren una base de evidencia más amplia y múltiples fuentes de información (11-13). En consecuencia, la revisión no formula una prohibición normativa del uso sumativo; señala que los estudios localizados todavía no aportan evidencia suficiente para justificar decisiones de alto impacto basadas exclusivamente en una salida automatizada.

4.6. Calidad pedagógica, validez y calibración del uso

La literatura sobre retroalimentación aporta criterios pedagógicos para interpretar las salidas generadas por IA. Hattie y Timperley señalaron que el efecto del retroalimentación depende de su contenido, nivel y forma de entrega, mientras van de Ridder et al. lo definieron en educación clínica como información específica que compara el desempeño observado con un estándar con intención

de mejorarlo (6,7). En simulación, el debriefing estructurado puede transformar la experiencia en aprendizaje (41,42). Estos marcos, incorporados desde la introducción, permiten distinguir tres preguntas que no deben confundirse: si el texto generado es correcto, si el retroalimentación es pedagógicamente accionable y si el juicio es válido para el uso evaluativo propuesto.

Desde una perspectiva de evaluación, la utilidad de un sistema automatizado depende del uso que se pretenda dar a sus juicios. Los criterios de Ottawa incluyen validez o coherencia, reproducibilidad, equivalencia, factibilidad, efecto educativo y aceptabilidad (11). El enfoque argumentativo de Kane subraya que se valida la interpretación y el uso de un resultado, no el puntaje en abstracto (12). La evaluación programática añade que las decisiones de progresión se fortalecen al integrar múltiples puntos de información y juicio experto (13). En consecuencia, la concordancia con un evaluador humano es una fuente de evidencia de validez, pero no agota la justificación necesaria para usos sumativos.

Tabla 4. Riesgos operativos de la retroalimentación automatizada y salvaguardas sugeridas.

Riesgo	Manifestación observada/posible	Salvaguarda	Indicador de auditoría	Base de la salvaguarda
Feedback inexacto	Omisiones, errores factuales, sobrepuntuación o juicio incorrecto de adherencia (19,26,35).	Rúbrica explícita + validación con docentes antes del uso.	% de acuerdo con evaluadores humanos por categoría.	Empírica + prudencial
Variabilidad	Evaluaciones diferentes ante el mismo texto o ejecuciones repetidas (19,26,35).	Versionado de modelo, temperatura controlada cuando sea posible y pruebas de repetibilidad.	Coeficiente de acuerdo/reproducibilidad en casos ancla.	Empírica + técnica
Exceso de positividad	Subestimación de errores o debilidades relevantes; tendencia a puntuaciones altas sin criterios estrictos (19,35).	Instrucciones para identificar fortalezas y brechas; umbrales de escalamiento.	Proporción de errores críticos omitidos.	Empírica + prudencial
Sesgo comunicativo	Penalización potencial de estilos lingüísticos/culturales distintos.	Pruebas con grupos diversos y revisión de equidad.	Diferencias sistemáticas por subgrupos o variante lingüística.	Prudencial/ética
Privacidad	Carga de conversaciones o datos sensibles en servicios externos.	Minimización de datos, acuerdos institucionales y plataformas aprobadas.	Incidentes, cumplimiento de políticas y trazabilidad del tratamiento de datos.	Prudencial/ética
Uso sumativo prematuro	Decisiones académicas basadas en una salida no validada.	Supervisión humana y confirmación independiente para alto impacto.	% de decisiones sumativas confirmadas por evaluador humano.	Prudencial/ validez

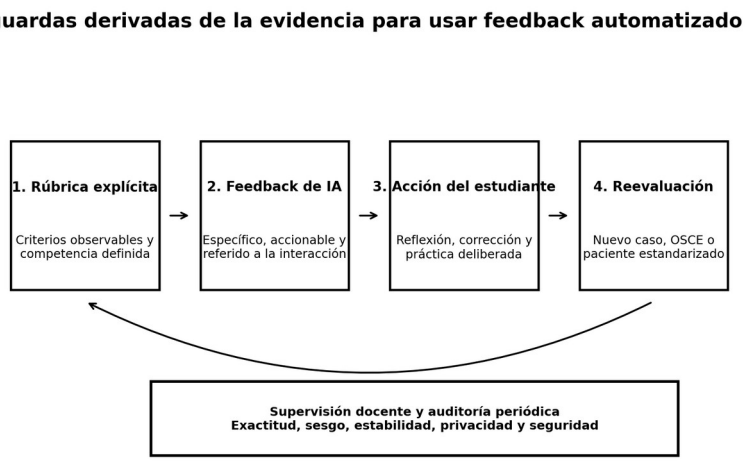

Figura 3. Salvaguardas derivadas de la evidencia para integrar retroalimentación automatizada en simulación clínica. La figura sintetiza principios de uso formativo derivados de los estudios revisados y de marcos de práctica deliberada, carga cognitiva y gobernanza ética; no representa un instrumento validado.

Esta interpretación coincide con síntesis contemporáneas. Feigerlova et al. encontraron evidencia heterogénea sobre resultados educativos de intervenciones con IA y escasa evaluación de transferencia al entorno clínico (43). Una revisión de alcance específica sobre retroalimentación generada por LLM en pregrado identificó predominio de resultados intermedios y exactitud variable (44). Además, revisiones recientes de pacientes virtuales con IA convergen en que los beneficios potenciales se concentran en práctica estructurada, entrevista, comunicación y razonamiento, mientras persisten incertidumbres sobre autenticidad, estandarización de outcomes y transferencia (14-17). Estudios experimentales adicionales muestran que la retroalimentación de ChatGPT puede aproximarse al retroalimentación experta en algunas tareas, pero no de forma uniforme, y que simples advertencias sobre errores no necesariamente modifican la conducta del estudiante (45-47).

4.7. Limitaciones de la revisión

Esta revisión tiene limitaciones importantes. Primero, aunque la identificación incluyó PubMed/MEDLINE, ERIC y OpenAlex y se contrastó la cobertura con revisiones multibase, no se realizó una búsqueda sistemática directa en todas las bases de suscripción; persiste riesgo de omisión. Segundo, el registro histórico de búsqueda no conservó un recuento exportable completo por fuente, duplicados y exclusiones, por lo que no fue posible reconstruir retrospectivamente un flujo cuantitativo auditable sin introducir cifras no verificables. Tercero, aunque la selección inicial y la extracción fueron realizadas por la autora principal y verificadas por una segunda autora, el proceso no equivalió a cribado independiente por duplicado desde la identificación. Cuarto, la heterogeneidad de competencias, diseños y métricas impidió una síntesis cuantitativa. Finalmente, la rápida obsolescencia tecnológica limita la transportabilidad temporal de los hallazgos: el rendimiento puede depender de la versión concreta del modelo, cambios de proveedor, prompts, parámetros de inferencia y actualizaciones posteriores a la publicación. En consecuencia, resultados

obtenidos con una versión específica no deben extrapolarse automáticamente a versiones futuras o a otros sistemas.

Por estas razones, el manuscrito debe interpretarse como una revisión crítica integrativa multifuente y no como una estimación exhaustiva del estado de la evidencia. Sus conclusiones se restringen a patrones observados en los estudios localizados y contrastados con síntesis recientes. Una futura revisión sistemática de efectividad debería incorporar búsqueda directa en múltiples bases de datos de suscripción, protocolo previo, cribado independiente por duplicado y, cuando exista homogeneidad suficiente, metaanálisis. También sería útil acordar un conjunto mínimo de outcomes que incluya exactitud frente a expertos, reproducibilidad, cambio de desempeño, transferencia a evaluación independiente, retención, carga cognitiva, equidad y costos.

5. Conclusiones

- Entre los 17 estudios localizados, se observaron resultados favorables en algunas tareas formativas y clínicas estructuradas, pero la magnitud y consistencia de esos resultados variaron según diseño, dominio y comparador.
- El rendimiento no fue uniforme. Se describieron sobrepuntuación, errores de especificidad, omisiones, variabilidad entre ejecuciones y menor concordancia en dominios complejos; por ello, fluidez lingüística y validez evaluativa no deben considerarse equivalentes.
- La evidencia de mayor peso interpretativo procede de ensayos aleatorizados y estudios con OSCE o pacientes estandarizados independientes; aun así, las muestras suelen ser pequeñas o moderadas, el seguimiento es corto y no se dispone de evidencia suficiente sobre retención o desempeño con pacientes reales.
- Los marcos de retroalimentación, práctica deliberada, carga cognitiva, validez y evaluación programática permiten interpretar por qué la utilidad de la IA depende de la tarea y del uso previsto. La evidencia actual respalda con mayor claridad aplicaciones formativas que decisiones sumativas autónomas de alto impacto.

Financiación: No ha habido financiación.

Declaración de conflicto de interés: Las autoras declaran no tener ningún conflicto de interés.

Contribuciones de las autoras: AMTM: conceptualización, metodología, investigación, búsqueda y selección inicial de la literatura, extracción y análisis de datos, visualización, redacción del borrador original, revisión y edición, y aprobación de la versión final. ADGC: metodología, validación, verificación independiente de la elegibilidad de los estudios incluidos y de la extracción de datos, análisis crítico, revisión y edición del manuscrito, y aprobación de la versión final.

Declaración sobre uso de inteligencia artificial: Se utilizó una herramienta de inteligencia artificial generativa como apoyo auxiliar para la estructuración inicial del manuscrito, organización de términos de búsqueda y revisión lingüística. Las autoras verificaron las fuentes bibliográficas, revisaron críticamente las síntesis y asumen la responsabilidad íntegra del contenido, interpretación y versión final del manuscrito. No se utilizaron datos confidenciales de participantes ni historias clínicas identificables.

6. Referencias

1. Fierro Manrique YV, Arteaga Losada MS, Anduquia Manzano C, Escalante Ochoa AM, Novoa Álvarez RA, Charry Cuellar JD. La simulación de alta fidelidad clínica con actuación y la simulación virtual en metaverso como estrategias para la formación de futuros médicos. *Rev Esp Edu Med.* **2026**, 5, 721571. https://revistas.um.es/edumed/article/view/721571. https://doi.org/10.6018/edumed.721571

2. Bonilla Mejia JJ, Cortés Fuenzalida TD, Polanco Aliaga DH, Martínez Carrillo C, Herrera Alcaíno AA. Inteligencia artificial en la formación clínica práctica de estudiantes de medicina de pregrado: una revisión de alcance de aplicaciones, resultados y brechas. *Rev Esp Edu Med.* **2026**, 7(3), 710071. https://doi.org/10.6018/edumed.710071

3. Thind BS, Javidi D, Schwartz LM. Artificial intelligence in undergraduate medical education clinical skills curricula: a scoping review of implementations since 2022. *Front Digit Health.* **2026**, 8, 1830254. https://doi.org/10.3389/fdgth.2026.1830254
4. Jiang J, Ye MZ, Kwok TTO, Wong JYH. GenAI-Supported Virtual Patients in Health Care Education: Systematic Review. *J Med Internet Res.* **2026**, 28, e82756. https://doi.org/10.2196/82756
5. Loubbairi S, El Moussaoui Y, Lahlou L, Chakri I, Nassik H. The impact of artificial intelligence-driven simulation on the development of non-technical skills in medical education: a systematic review. *J Educ Eval Health Prof.* **2025**, 22, 37. https://doi.org/10.3352/jeehp.2025.22.37
6. Hattie J, Timperley H. The Power of Feedback. *Rev Educ Res.* **2007**, 77(1), 81-112. https://doi.org/10.3102/003465430298487
7. van de Ridder JMM, Stokking KM, McGaghie WC, ten Cate OTJ. *What is feedback in clinical education? Med Educ.* **2008**, 42(2), 189-197. https://doi.org/10.1111/j.1365-2923.2007.02973.x
8. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med.* **2011**, 86(6), 706-711. https://doi.org/10.1097/ACM.0b013e318217e119
9. Young JQ, Van Merrienboer J, Durning S, Ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86. *Med Teach.* **2014**, 36(5), 371-384. https://doi.org/10.3109/0142159X.2014.889290
10. Fraser KL, Ayres P, Sweller J. Cognitive Load Theory for the Design of Medical Simulations. *Simul Healthc.* **2015**, 10(5), 295-307. https://doi.org/10.1097/SIH.0000000000000097
11. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* **2011**, 33(3), 206-214. https://doi.org/10.3109/0142159X.2011.551559
12. Kane MT. Validating the Interpretations and Uses of Test Scores. *J Educ Meas.* **2013**, 50(1), 1-73. https://doi.org/10.1111/jedm.12000
13. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, van Tartwijk J. A model for programmatic assessment fit for purpose. *Med Teach.* **2012**, 34(3), 205-214. https://doi.org/10.3109/0142159X.2012.652239
14. Bowers P, Graydon K, Ryan T, Lau JH, Tomlin D. Artificial intelligence-driven virtual patients for communication skill development in healthcare students: A scoping review. *Australas J Educ Technol.* **2024**, 40(3), 39-57. https://doi.org/10.14742/ajet.9307
15. Gilbert V, Philip P, Llorca PM, Samalin L. Use of artificial intelligence-enhanced virtual patients in educational approaches to medical interview training: a systematic review. *BMC Med Educ.* **2026**, 26, 588. https://doi.org/10.1186/s12909-026-08804-9
16. Fang EKM, Boon Y, Ho JY, Tey N, Low MJW, et al. Artificial intelligence generated virtual patients in health professions education: a scoping review. BMC Med Educ. 2026. https://doi.org/10.1186/s12909-026-10028-w
17. Mofidi SA, Khoshgoftar Z. Effectiveness of AI-enhanced virtual patients in clinical reasoning training of healthcare professionals and students: a systematic review. BMC Med Educ. 2026. https://doi.org/10.1186/s12909-026-10242-6
18. Brügge E, Ricchizzi S, Arenbeck M, Keller MN, Schur L, Stummer W, Holling M, Lu MH, Darici D. 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, 1391. https://doi.org/10.1186/s12909-024-06399-7
19. Wang C, Li S, Lin N, Zhang X, Han Y, Wang X, Liu D, Tan X, Pu D, Li K, Qian G, Yin R. Application of Large Language Models in Medical Training Evaluation-Using ChatGPT as a Standardized Patient: Multimetric Assessment. *J Med Internet Res.* **2025**, 27, e59435. https://doi.org/10.2196/59435

20. Snedegar R, Unger K, Craig JF, Kozlowski L, Carpenter D, Pyles E, Williamson J, Bodkins E, Williams D. Ambient Listening Devices as a Feedback Tool in the Simulated Learning Environment. *Cureus.* **2026**, 18(2), e103222. https://doi.org/10.7759/cureus.103222
21. Nolan EJ, Burke HB. Large Language Model Evaluation and Feedback of Transcribed Simulated Physician-Patient Verbal Interactions. *Front Artif Intell.* **2026**, 9, 1936749. https://doi.org/10.3389/frai.2026.1936749
22. Maicher KR, Zimmerman L, Wilcox B, et al. Using virtual standardized patients to accurately assess information gathering skills in medical students. *Med Teach.* **2019**, 41(9), 1053-1059. https://doi.org/10.1080/0142159X.2019.1616683
23. Holderried F, Stegemann-Philipps C, Herrmann-Werner A, Festl-Wietek T, Holderried M, Eickhoff C, Mahling M. 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
24. Yamamoto A, Koda M, Ogawa H, Miyoshi T, Maeda Y, Otsuka F, Ino H. Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial. *JMIR Med Educ.* **2024**, 10, e58753. https://doi.org/10.2196/58753
25. Scherr R, Spina A, Dao A, Andalib S, Halaseh FF, Blair S, Wiechmann W, Rivera R. Novel Evaluation Metric and Quantified Performance of ChatGPT-4 Patient Management Simulations for Early Clinical Education: Experimental Study. *JMIR Form Res.* **2025**, 9, e66478. https://doi.org/10.2196/66478
26. Weisman D, Sugarman A, Huang YM, Gelberg L, Ganz PA, Comulada WS. Development of a GPT-4-Powered Virtual Simulated Patient and Communication Training Platform for Medical Students to Practice Discussing Abnormal Mammogram Results With Patients: Multiphase Study. *JMIR Form Res.* **2025**, 9, e65670. https://doi.org/10.2196/65670
27. Laverde N, Grévisse C, Jaramillo S, Manrique R. Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. *Comput Struct Biotechnol J.* **2025**, 27, 2481-2491. https://doi.org/10.1016/j.csbj.2025.05.025
28. Suárez-García RX, Chavez-Castañeda Q, Orrico-Pérez R, et al. DIALOGUE: A Generative AI-Based Pre-Post Simulation Study to Enhance Diagnostic Communication in Medical Students Through Virtual Type 2 Diabetes Scenarios. *Eur J Investig Health Psychol Educ.* **2025**, 15(8), 152. https://doi.org/10.3390/ejihpe15080152
29. Jacobs C, Johnson H, Tan N, Brownlie K, Joiner R, Thompson T. Application of AI Communication Training Tools in Medical Undergraduate Education: Mixed Methods Feasibility Study Within a Primary Care Context. *JMIR Med Educ.* **2025**, 11, e70766. https://doi.org/10.2196/70766
30. Luo MJ, Bi S, Pang J, 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
31. Byun J, Kim H, Lim J, Choi J, Ahn S. Enhancing history-taking education through GPT-4-based virtual patients and automated assessment: a study of medical student perceptions. *Korean J Med Educ.* **2026**, 38(1), 64-73. https://doi.org/10.3946/kjme.2025.108
32. Borg A, Schiött J, Ivegren W, 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. https://doi.org/10.2196/90368
33. Sindhvananda K, Ruengwattanachot K, Leksuwankun S, et al. AI-powered simulated patients with automated feedback for enhancing headache history-taking skills: a convergent mixed-methods study. *BMC Med Educ.* **2026**, 26, 1151. https://doi.org/10.1186/s12909-026-09511-1
34. Chan BS, Marjot J, Dodds T, et al. Enhancing objective structured clinical examination performance through an artificial intelligence virtual patient: a proof-of-concept study. *Proc (Bayl Univ Med Cent).* **2026**, 1-4. https://doi.org/10.1080/08998280.2026.2699604

35. Cook DA, Overgaard J, Pankratz VS, Del Fiol G, Aakre CA. Virtual Patients Using Large Language Models: Scalable, Contextualized Simulation of Clinician-Patient Dialogue With Feedback. *J Med Internet Res.* **2025**, 27, e68486. https://doi.org/10.2196/68486
36. Lee J, Kim H, Kim KH, Jung D, Jowsey T, Webster CS. Effective virtual patient simulators for medical communication training: a systematic review. *Med Educ.* **2020**, 54(9), 786-795. https://doi.org/10.1111/medu.14152
37. Kononowicz AA, Woodham LA, Edelbring S, et al. Virtual Patient Simulations in Health Professions Education: Systematic Review and Meta-Analysis by the Digital Health Education Collaboration. *J Med Internet Res.* **2019**, 21(7), e14676. https://doi.org/10.2196/14676
38. Durning SJ, Artino AR Jr. Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide No. 52. *Med Teach.* **2011**, 33(3), 188-199. https://doi.org/10.3109/0142159X.2011.550965
39. Hippe DS, Umoren RA, McGee A, Bucher SL, Bresnahan BW. A targeted systematic review of cost analyses for implementation of simulation-based education in healthcare. *SAGE Open Med.* **2020**, 8, 2050312120913451. https://doi.org/10.1177/2050312120913451
40. Masters K. Ethical use of Artificial Intelligence in Health Professions Education: AMEE Guide No. 158. *Med Teach.* **2023**, 45(6), 574-584. https://doi.org/10.1080/0142159X.2023.2186203
41. Cantillon P, Sargeant J. Giving feedback in clinical settings. *BMJ.* **2008**, 337, a1961. https://doi.org/10.1136/bmj.a1961
42. Cheng A, Eppich W, Grant V, Sherbino J, Zendejas B, Cook DA. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. *Med Educ.* **2014**, 48(7), 657-666. https://doi.org/10.1111/medu.12432
43. 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
44. Kıyak YS, İş-Kara T, Emekli E. Applications and Outcomes of Large-Language-Model-Generated Feedback in Undergraduate Medical Education: A Scoping Review. *Med Sci Educ.* **2026**, 36(1), 81-99. https://doi.org/10.1007/s40670-025-02621-3
45. Çiçek FE, Ülker M, Özer M, Kıyak YS. ChatGPT versus expert feedback on clinical reasoning questions and their effect on learning: a randomized controlled trial. *Postgrad Med J.* **2025**, 101(1199), 458-463. https://doi.org/10.1093/postmj/qgae170
46. 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
47. Kıyak YS, Coşkun Ö, Budakoğlu İİ. 'ChatGPT can make mistakes' warnings fail: A randomized controlled trial. *Med Educ.* **2026**, 60(2), 138-142. https://doi.org/10.1111/medu.70056
54. Elizondo-García J. Vibe-coding como competencia docente en el uso de Inteligencia Artificial para la Educación Médica: una revisión de alcance. *Rev Esp Edu Med.* **2026**, 7(2). https://revistas.um.es/edumed/article/view/705121, https://doi.org/10.6018/edumed.705121
53. Kaya AB, Emekli E, Kiyak YS. Mapeo de aplicaciones y resultados de casos generados por modelos de lenguaje amplio en la educación de profesiones de la salud: una revisión de alcance. *Rev Esp Edu Med.* **2026**, 7(1). https://revistas.um.es/edumed/article/view/691691, https://doi.org/10.6018/edumed.691691
52. Jiménez Salcedo SC, Granda Cruz CA. Uso de ChatGPT como herramienta pedagógica en la educación médica: revisión sistemática de la literatura (2020-2025). *Rev Esp Edu Med.* **2026**, 7(4). https://doi.org/10.6018/edumed.716681
51. Canabal Berlanga A, Sánchez Giralt JA, Abad Santamaría B. Actualidad en la educación médica con la inteligencia artificial generativa. *Rev Esp Edu Med.* **2025**, 6(6). https://doi.org/10.6018/edumed.684461

50. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Educ Inf.* **2018**, 34(4), 285-291. https://doi.org/10.3233/EFI-180221
49. Snyder H. Literature review as a research methodology: An overview and guidelines. *J Bus Res.* **2019**, 104, 333-339. https://doi.org/10.1016/j.jbusres.2019.07.039
48. Whittemore R, Knafl K. The integrative review: updated methodology. *J Adv Nurs.* **2005**, 52(5), 546-553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

© 2026 University of Murcia. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 Spain License (CC BY-NC-ND) (http://creativecommons.org/licenses/by/4.0/)

7. Anexo

Anexo 1. Estrategias de búsqueda reproducibles utilizadas en la revisión crítica integrativa multifuente

PubMed/MEDLINE (actualizada el 3 de septiembre de 2026): (("medical student"[Title/Abstract] OR "medical students"[Title/Abstract] OR "medical education"[Title/Abstract]) AND ("virtual patient"[Title/Abstract] OR "simulated patient"[Title/Abstract] OR "standardized patient"[Title/Abstract] OR "clinical simulation"[Title/Abstract] OR simulation[Title/Abstract]) AND ("artificial intelligence"[Title/Abstract] OR "generative AI"[Title/Abstract] OR GPT[Title/Abstract] OR "large language model"[Title/Abstract]) AND (feedback[Title/Abstract] OR "automated assessment"[Title/Abstract] OR scoring[Title/Abstract] OR evaluation[Title/Abstract])).

ERIC (actualizada el 3 de septiembre de 2026): ("medical students" OR "medical education") AND ("virtual patient" OR "simulated patient" OR "standardized patient" OR "clinical simulation") AND ("artificial intelligence" OR "large language model" OR GPT OR "generative AI") AND (feedback OR assessment OR scoring OR evaluation). La búsqueda se ejecutó sobre los campos bibliográficos disponibles en la interfaz, sin filtro de texto completo. OpenAlex (actualizada el 3 de septiembre de 2026): búsqueda por términos en título/resumen/conceptos mediante combinaciones de "AI-generated feedback", "automated feedback", "medical education", "virtual patient", "standardized patient", "clinical simulation", "large language model", GPT, assessment y scoring; los resultados se verificaron por título y DOI y se siguieron las relaciones de citación hacia atrás y hacia adelante. Se consideraron publicaciones en inglés o español y no se aplicó filtro por disponibilidad de texto completo. El procedimiento se complementó con rastreo de referencias y contraste con revisiones multibase recientes (14-17).