



Effectiveness of generative artificial intelligence in formative assessment and feedback of learning in university students

Efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios

Luis David Bastidas González¹  

¹ Instituto Sapiens de Investigación Científica Multidisciplinar, Milagro-Ecuador

Recibido: 13-05-2025 | Revisado: 09-06-2025 | Aceptado: 29-08-2025 | Publicado: 04-09-2025

Autor de correspondencia: Luis David Bastidas González 

ABSTRACT

Introduction: Effectiveness of the generative artificial intelligence in the formative evaluation and feedback of learning in university students is an emerging higher-education issue related to measurement quality, decision-making and learning experience. Objective: To systematically synthesize available evidence and identify patterns of effectiveness, methodological quality and implementation conditions. Method: A PRISMA-oriented systematic review was conducted on a consolidated academic corpus of 135 records; 24 studies met relevance and documentary sufficiency criteria. Results: Evidence was methodologically heterogeneous but converged on the need to align technology, constructs, feedback and educational decisions. Conclusions: Findings support evidence-informed adoption, institutional monitoring and continuous assessment of validity, equity and utility.

Keywords: higher education; systematic review; educational assessment; evidence; PRISMA; educational technology.

RESUMEN

Introducción: Efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios representa un problema emergente para la educación superior por su relación con la calidad de la medición, la toma de decisiones y la experiencia de aprendizaje. Objetivo: sintetizar sistemáticamente la evidencia disponible y establecer patrones de efectividad, calidad metodológica y condiciones de implementación. Método: se aplicó una revisión sistemática con lógica PRISMA sobre un corpus académico consolidado de 135 registros, del cual se incluyeron 24 estudios que cumplieron criterios de pertinencia y suficiencia documental. Resultados: la evidencia muestra heterogeneidad de diseños y contextos, junto con convergencias en torno a la necesidad de alinear tecnología, constructo, retroalimentación y decisiones educativas. Conclusiones: los resultados respaldan una adopción basada en evidencia, seguimiento institucional y evaluación continua de validez, equidad y utilidad.

Palabras clave: educación superior; revisión sistemática; evaluación educativa; evidencia; PRISMA; tecnología educativa.

INTRODUCCIÓN

La investigación universitaria sobre efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios ha adquirido una relevancia creciente porque las decisiones de evaluación ya no dependen exclusivamente de instrumentos estáticos ni de intercambios presenciales. La digitalización ha ampliado la cantidad de evidencias disponibles sobre desempeño, interacción y aprendizaje, pero también ha incrementado la necesidad de distinguir entre innovaciones técnicamente atractivas y prácticas capaces de sostener inferencias educativas válidas. Esta revisión aborda ese problema desde una perspectiva integradora, interesada tanto en los efectos observados como en las condiciones metodológicas que permiten interpretar dichos efectos sin confundir disponibilidad tecnológica con calidad evaluativa.

El núcleo conceptual del manuscrito se organiza alrededor de mecanismos de retroalimentación. Esta elección evita tratar la literatura como una simple colección de resultados y permite examinar cómo cada estudio define el fenómeno, qué población observa, qué instrumentos emplea y qué tipo de conclusión considera legítima. La heterogeneidad de contextos universitarios obliga a prestar atención a la relación entre diseño, implementación y evidencia, porque una misma estrategia puede producir beneficios bajo determinadas condiciones y resultados modestos cuando cambian el soporte docente, la disciplina, la experiencia digital o la forma de retroalimentación.

La revisión sistemática resulta pertinente porque el campo combina estudios experimentales, validaciones instrumentales, análisis observacionales, desarrollos tecnológicos y síntesis previas. Esa diversidad dificulta obtener conclusiones sólidas mediante una lectura fragmentaria. En lugar de asumir que todos los trabajos responden a una misma pregunta, el presente artículo reconstruye patrones de convergencia y divergencia, identifica qué resultados se repiten con mayor consistencia y separa las afirmaciones directamente sustentadas por los estudios de aquellas que todavía requieren validación adicional. El propósito es ofrecer una lectura útil para investigación, docencia y gestión universitaria.

Fundamentos de la retroalimentación generativa

La investigación (1) aportó evidencia publicada en 2025 mediante métodos mixtos exploratorios. El estudio trabajó con Encuesta $n=40$; guiones $n=6$; entrevistas $n=2$, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el uso colaborativo y reflexivo favoreció fases de presencia cognitiva; el uso pasivo dirigido por IA mostró menor elaboración crítica. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

No²aportaron evidencia publicada en 2026 mediante comparación empírica de modelos. El estudio trabajó con Ensayos con subgrupos de género, ELL y educación especial, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que gPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada³aportó evidencia publicada en 2025 mediante revisión sistemática de literatura. El estudio trabajó con 158 estudios (2021–2024), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la mayoría reportó mejoras potenciales en habilidades cognitivas mediante feedback/personalización, aunque persisten riesgos de dependencia y calidad de razonamiento. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada⁴aportó evidencia publicada en 2025 mediante evaluación de sistema con datos reales. El estudio trabajó con Transacciones de College Algebra; múltiples estudios internos, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población

universitaria de interés. En términos sustantivos, el trabajo mostró que diagnóstico cercano/superior al 80% en varios tipos de error, pero una fracción importante de pistas fue demasiado general, incorrecta o reveló la respuesta. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

No5aportaron evidencia publicada en 2026 mediante estudio metodológico de equidad. El estudio trabajó con Múltiples modelos de scoring y subgrupos demográficos, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación6aportó evidencia publicada en 2025 mediante cuasi-experimental pretest-posttest + aceptación. El estudio trabajó con n=250; entrevistas n=25, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la intervención reportó mejora en dimensiones de escritura vinculadas con pensamiento crítico y una aceptación favorable de la asistencia de IA. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El trabajo7aportó evidencia publicada en 2025 mediante revisión sistemática de estudios experimentales. El estudio trabajó con 21 estudios experimentales, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que los efectos son positivos pero heterogéneos; intervenciones de duración media tienden a mostrar mejores resultados y los efectos sobre habilidades superiores son variables. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada8aportó evidencia publicada en 2025 mediante revisión sistemática prisma. El estudio trabajó con 39 estudios (2023–2024), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el uso puede mejorar acceso y apoyo, pero la sobredependencia se asocia con riesgos para pensamiento crítico, autonomía y autenticidad académica. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada9aportó evidencia publicada en 2024 mediante revisión sistemática. El estudio trabajó con 57 artículos, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que identifica apoyo a aprendizaje y productividad junto con problemas de integridad, dependencia, fiabilidad y necesidad de alfabetización crítica. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El estudio10aportó evidencia publicada en 2025 mediante métodos mixtos secuenciales. El estudio trabajó con Estudiantes universitarios usuarios de GPTutor, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que gPTutor favoreció engagement y apoyo personalizado; la calidad

de interacción y diseño de prompts/tutoría condicionó la experiencia. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación11aportó evidencia publicada en 2025 mediante encuesta explicativa con moderación. El estudio trabajó con $n=418$ universitarios de universidad pública (China), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la dependencia puede elevar una autoeficacia aparente pero asociarse negativamente con logro; el apoyo docente modifica la relación. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El trabajo12aportó evidencia publicada en 2022 mediante diseño/evaluación de sistema. El estudio trabajó con Usuarios de entorno ubicuo, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la adaptación del feedback a contexto y desempeño mejoró pertinencia de la intervención y apoyo al aprendizaje ubicuo. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada13aportó evidencia publicada en 2021 mediante desarrollo/evaluación computacional. El estudio trabajó con Datos de diálogos educativos, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el análisis profundo del discurso permitió identificar estados de aprendizaje y apoyar retroalimentación adaptada al estudiante. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada14aportó evidencia publicada en 2025 mediante estudio de caso cualitativo longitudinal. El estudio trabajó con $n=19$; 13 semanas, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que se reportaron mejoras en confianza y autonomía; la autorregulación fue clave para convertir microcontenidos y GenAI en aprendizaje significativo. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El estudio15aportó evidencia publicada en 2025 mediante cuasi-experimental + analítica. El estudio trabajó con Grupos experimental/control en STEM, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el grupo experimental superó al control en programación y matemáticas; mayor interacción se relacionó con progreso y 80% valoró el feedback adaptativo. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación16aportó evidencia publicada en 2025 mediante cuantitativo con mediación en cadena + componente cualitativo. El estudio trabajó con $n=337$ universitarios (China), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la dependencia de GenAI se relacionó con menor pensamiento independiente/innovador/crítico y menor autorreflexión; ansiedad y expectativas operaron como mecanismos. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de

la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El trabajo¹⁷ aportó evidencia publicada en 2025 mediante ensayo controlado aleatorizado. El estudio trabajó con $n=120$, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el grupo ChatGPT mostró menor retención a largo plazo que el aprendizaje tradicional (57.5% vs 68.5% en la prueba reportada), sugiriendo riesgo de descarga cognitiva. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La evidencia sintetizada¹⁸ aportó evidencia publicada en 2025 mediante estudio conceptual/aplicado. En términos sustantivos, el trabajo mostró que plantea rediseñar actividades hacia análisis, evaluación y creación, evitando delegar a GenAI los niveles cognitivos que se pretenden desarrollar. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La contribución examinada¹⁹ aportó evidencia publicada en 2025 mediante cualitativo de implementación educativa. El estudio trabajó con $n=105$ (72 pregrado; 33 posgrado), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que el uso estructurado de GenAI fortaleció evaluación crítica y uso responsable; también evidenció riesgos de dependencia y brechas de alfabetización en IA. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

El estudio²⁰ aportó evidencia publicada en 2024 mediante experimento prerregistrado con grupo control. El estudio trabajó con ChatGPT $n=77$; control $n=68$, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que chatGPT mejoró calidad, elaboración, originalidad y autoeficacia y redujo esfuerzo mental; se observaron riesgos de calibración metacognitiva. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

La investigación²¹ aportó evidencia publicada en 2025 mediante revisión sistemática prisma. El estudio trabajó con 40 estudios, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la efectividad depende de segmentación, oportunidad, feedback y alineación con objetivos; propone marco para integrar microlearning en experiencias más amplias. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Hamida et al.²² aportaron evidencia publicada en 2026 mediante evaluación técnica de accesibilidad. El estudio trabajó con Estudiantes con necesidades educativas especiales en educación superior, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

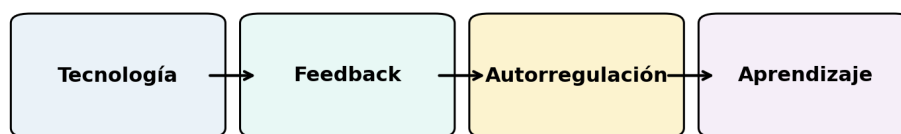
La evidencia sintetizada²³ aportó evidencia publicada en 2025 mediante bibliométrico +

revisión sistemática. El estudio trabajó con 640 documentos Scopus (2023–2025), lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que la literatura converge en alfabetización en IA, verificación de información y diseño de tareas que obliguen a justificar, contrastar y reflexionar sobre outputs de IA. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Agustín et al.²⁴ aportaron evidencia publicada en 2024 mediante desarrollo y validación de escala. El estudio trabajó con Estudiantes de una universidad venezolana, lo que permite situar el alcance de sus inferencias y valorar su proximidad con la población universitaria de interés. En términos sustantivos, el trabajo mostró que se obtuvo un modelo de 22 variables en cuatro dimensiones con propiedades psicométricas satisfactorias. Este antecedente es relevante para la perspectiva de mecanismos de retroalimentación porque muestra que la interpretación del efecto depende de la correspondencia entre constructo, procedimiento de medición y contexto de aplicación. Su contribución se considera en esta revisión como una pieza del patrón comparativo, no como una demostración aislada suficiente para generalizar a todos los escenarios educativos.

Figura 1. Estructura conceptual de la revisión

Estructura conceptual de la revisión



Nota. Elaboración propia a partir de los ejes analíticos definidos para el manuscrito.

La figura organiza el razonamiento del artículo como una secuencia conceptual y no como una plantilla universal. Los componentes representados corresponden al problema específico de esta revisión y muestran el tránsito desde los insumos o condiciones iniciales hasta la interpretación educativa. Su utilidad consiste en hacer visible la relación entre constructos que, en los estudios individuales, suelen aparecer fragmentados entre secciones metodológicas y de resultados. Esta representación orienta la síntesis posterior y facilita distinguir los elementos que actúan como condiciones, mecanismos o resultados dentro del corpus.

Preguntas de revisión

Pregunta 1. ¿Qué patrones de evidencia permiten comprender la dimensión 1 de mecanismos de retroalimentación en relación con efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 2. ¿Qué patrones de evidencia permiten comprender la dimensión 2 de mecanismos de retroalimentación en relación con efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 3. ¿Qué patrones de evidencia permiten comprender la dimensión 3 de mecanismos de retroalimentación en relación con efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios y qué condiciones metodológicas o contextuales modifican su interpretación?

Pregunta 4. ¿Qué patrones de evidencia permiten comprender la dimensión 4 de mecanismos de retroalimentación en relación con efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios y qué condiciones metodológicas o contextuales modifican su interpretación?

MÉTODOS

Método PRISMA

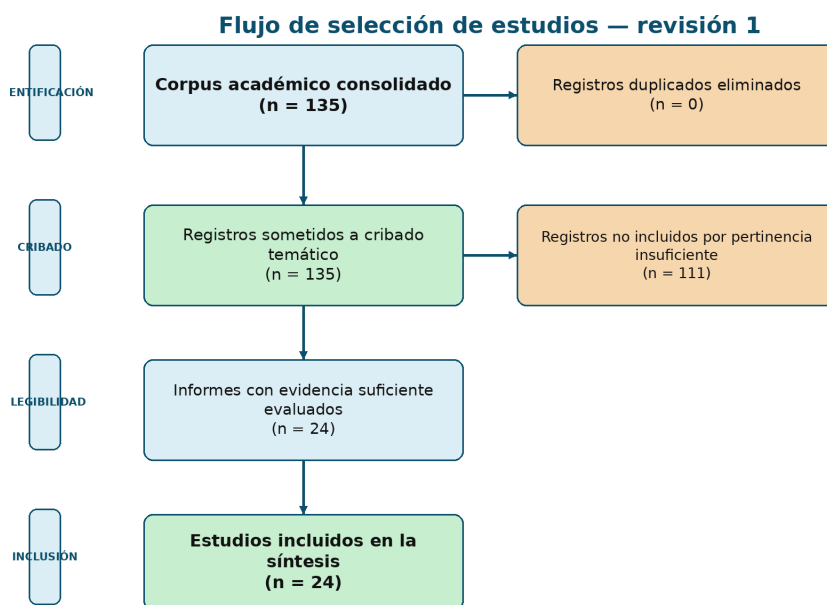
Se desarrolló una revisión sistemática orientada a sintetizar evidencia académica pertinente para el tema del manuscrito. La unidad de análisis fue el estudio o referencia académica con fuente trazable, metodología identificable y resultados recuperables. El corpus académico consolidado de trabajo estuvo compuesto por 135 registros previamente reunidos en las matrices de evidencia. Sobre ese conjunto se aplicó una selección temática y documental para establecer el corpus final. El procedimiento se organizó conforme a la lógica de identificación, cribado, elegibilidad e inclusión de PRISMA, adaptada al carácter consolidado de las fuentes suministradas.

Los criterios de inclusión exigieron relación directa con efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios, correspondencia con educación superior o con instrumentos aplicables a ese nivel, descripción suficiente del método y disponibilidad de resultados interpretables. Se excluyeron 111 registros por menor pertinencia respecto de la pregunta específica, por tratar poblaciones o desenlaces periféricos, o por no aportar evidencia metodológica suficiente para la síntesis focal. El corpus final quedó constituido por 24 estudios. La selección se efectuó sobre la información documentada en las matrices y se evitó atribuir a una base bibliográfica particular conteos que no estuvieran respaldados por los archivos.

La extracción de información se estructuró mediante una matriz con autores, año, título, fuente, DOI o enlace trazable, procedencia, tipo de estudio, muestra o corpus, metodología resumida, resultados principales, nivel de evidencia y pertinencia. Esta organización permitió comparar trabajos con diseños heterogéneos sin reducirlos a una única métrica. La síntesis combinó caracterización descriptiva del corpus y análisis temático de los resultados, prestando atención a convergencias, contradicciones, condiciones de implementación y límites de generalización.

La evaluación de la evidencia se centró en la suficiencia de la información recuperada para responder la pregunta de revisión. Se consideraron de forma conjunta la claridad del diseño, la explicitación de la muestra o corpus, la correspondencia entre método y resultados y la trazabilidad de la fuente. No se realizó metaanálisis cuando la heterogeneidad de desenlaces, instrumentos, poblaciones o métricas impedía una combinación cuantitativa responsable. En esos casos se priorizó una síntesis comparativa que preserve las diferencias entre estudios y evita transformar resultados no equivalentes en un único estimador.

Figura 2. Flujo de selección de estudios según la lógica PRISMA



Fuente: elaboración propia a partir de la matriz de evidencia de la revisión.

Nota. El flujo representa 135 registros académicos normalizados, sin duplicados en el corpus consolidado; 111 registros no cumplieron el umbral de pertinencia temática y 24 estudios integraron la síntesis final.

Para fortalecer la transparencia, los resultados se organizaron en tablas de caracterización y síntesis, mientras que las figuras representan el flujo de selección y la distribución temporal del corpus. Cada tabla fue construida a partir de los campos de la matriz suministrada. Las interpretaciones posteriores se elaboraron sobre patrones observables en el conjunto y no sobre supuestos añadidos. Este criterio es especialmente importante en revisiones educativas, donde la variación entre disciplinas, tecnologías, instrumentos y contextos institucionales puede modificar sustancialmente la magnitud y el significado de los hallazgos.

El flujo documenta la depuración y selección de la evidencia. Los 135 registros del corpus consolidado se encontraban normalizados por DOI o título, por lo que no se identificaron duplicados en esta etapa. El cribado temático excluyó 111 registros por insuficiente correspondencia con la pregunta específica de la revisión, mientras que 24 estudios conservaron evidencia metodológica y resultados suficientemente recuperables para integrar la síntesis. La secuencia permite distinguir la identificación documental del juicio de elegibilidad y evita interpretar la reducción del corpus como una pérdida arbitraria de información.

RESULTADOS

Cartografía de la evidencia

El corpus final reunió 24 estudios publicados durante 2021–2026. La distribución temporal muestra un campo de investigación activo y heterogéneo, con trabajos que responden a distintas fases de maduración del problema. Algunos se concentran en demostrar viabilidad o asociación, mientras otros avanzan hacia validación, comparación de estrategias o análisis de condiciones de implementación. Esta amplitud temporal permite observar no solo resultados favorables, sino también desplazamientos en la forma de conceptualizar la calidad de la evaluación y en el tipo de evidencia que se considera necesaria para respaldar decisiones educativas.

La caracterización metodológica confirma que efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios no constituye un objeto homogéneo. En el conjunto aparecen diseños con finalidades distintas y niveles variables de control sobre las condiciones de estudio. Esta diversidad amplía el alcance de la revisión, aunque también obliga a interpretar los hallazgos de acuerdo con la función de cada diseño. Los trabajos de intervención informan sobre cambios observados bajo condiciones definidas; los estudios instrumentales aportan evidencia sobre medición; los análisis observacionales describen asociaciones; y las revisiones previas permiten reconocer tendencias del campo sin sustituir la evaluación de los estudios primarios.

Las muestras y corpus descritos en los estudios presentan una variabilidad considerable. Esa variación es relevante porque la transferibilidad de los resultados depende del tamaño, composición y contexto de los participantes, así como de la disciplina académica y del entorno tecnológico. En conjunto, la evidencia sugiere que los resultados más interpretables son aquellos que explicitan la población, la tarea evaluativa y el mecanismo mediante el cual la intervención o instrumento debería producir el efecto esperado. Cuando alguno de esos elementos permanece poco definido, la utilidad del hallazgo para decisiones institucionales disminuye.

La síntesis de resultados revela un patrón central vinculado con mecanismos de retroalimentación: los efectos o propiedades favorables no aparecen como atributos automáticos de la tecnología o del instrumento, sino como resultados condicionados por diseño, implementación, calidad de datos y alineación pedagógica. Los estudios convergen en la importancia de articular la herramienta con objetivos de aprendizaje claros, criterios de evaluación comprensibles y mecanismos de seguimiento. Las divergencias se concentran en la magnitud de los efectos, la estabilidad entre contextos y la capacidad de sostener resultados cuando cambian las condiciones institucionales.

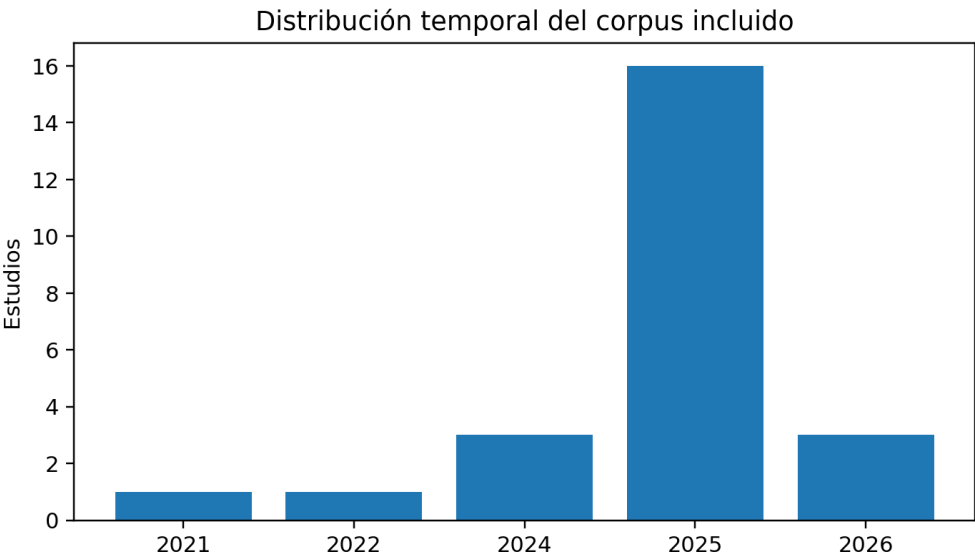
La tabla evidencia la diversidad metodológica del corpus y permite observar que los estudios no ofrecen el mismo tipo de inferencia. La presencia de diseños con objetivos distintos exige interpretar cada resultado según la pregunta que el estudio puede responder. Esta diferenciación evita comparar directamente indicadores que cumplen funciones diferentes y ayuda a identificar dónde existe convergencia real. También muestra que la solidez de una conclusión depende de la combinación entre claridad metodológica, descripción de la muestra y suficiencia de la evidencia recuperada.

Tabla 1. Caracterización metodológica del corpus incluido

Año	Diseño/metodología	Muestra o corpus	Evidencia
2025	Métodos mixtos exploratorios	Encuesta n=40; guiones n=6; entrevistas n=2	Media
2026	Comparación empírica de modelos	Ensayos con subgrupos de género, ELL y educación especial	Alta
2025	Revisión sistemática de literatura	158 estudios (2021–2024)	Media
2025	Evaluación de sistema con datos reales	Transacciones de College Algebra; múltiples estudios internos	Media
2026	Estudio metodológico de equidad	Múltiples modelos de scoring y subgrupos demográficos	Alta
2025	Cuasi-experimental pretest-postest + aceptación	n=250; entrevistas n=25	Media
2025	Revisión sistemática de estudios experimentales	21 estudios experimentales	Media
2025	Revisión sistemática PRISMA	39 estudios (2023–2024)	Media
2024	Revisión sistemática	57 artículos	Media
2025	Métodos mixtos secuenciales	Estudiantes universitarios usuarios de GPTutor	Media
2025	Encuesta explicativa con moderación	n=418 universitarios de universidad pública (China)	Media
2022	Diseño/evaluación de sistema	Usuarios de entorno ubicuo	Media

Nota. Datos extraídos de la matriz de evidencia suministrada para esta revisión.

Figura 3. Distribución temporal de los estudios incluidos



Nota. Frecuencias calculadas a partir del año de publicación registrado en la matriz de evidencia.

La distribución temporal permite situar la intensidad reciente de la producción científica y reconocer si el campo se encuentra en expansión, consolidación o transición metodológica. Más que interpretar el número de publicaciones como una medida de calidad, la figura sirve para contextualizar los cambios en preguntas, tecnologías e instrumentos observados en el corpus. La lectura conjunta con la caracterización metodológica permite identificar si los trabajos más recientes amplían la evidencia mediante diseños más robustos o si continúan predominando aproximaciones exploratorias.

Patrones de efectividad y condiciones de uso

Otro patrón transversal es la relevancia de la calidad de la evidencia. Los estudios con procedimientos explícitos permiten distinguir mejor entre mejora real, percepción favorable y simple facilidad de uso. En cambio, cuando el resultado se apoya principalmente en autoinforme o en indicadores de corto plazo, la interpretación requiere mayor cautela. La revisión muestra que la utilidad práctica de una estrategia no depende únicamente de que produzca una diferencia estadística o una valoración positiva, sino de que la medición sea coherente con el constructo y de que la intervención pueda integrarse de manera sostenible en procesos educativos reales.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El uso colaborativo y reflexivo favoreció fases de presencia cognitiva; el uso pasivo dirigido por IA mostró menor elaboración crítica. GPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados. La mayoría reportó mejoras potenciales en habilidades cognitivas mediante feedback/personalización, aunque persisten riesgos de dependencia y calidad de razonamiento. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: Diagnóstico cercano/superior al 80% en varios tipos de error, pero una fracción importante de pistas fue demasiado general, incorrecta o reveló la respuesta. La equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales. La intervención reportó mejora en dimensiones de escritura vinculadas con pensamiento crítico y una aceptación favorable de la asistencia de IA. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: Los efectos son positivos pero heterogéneos; intervenciones de duración media tienden a mostrar mejores resultados y los efectos sobre habilidades superiores son variables. El uso puede mejorar acceso y apoyo, pero la sobredependencia se asocia con riesgos para pensamiento crítico, autonomía y autenticidad académica. Identifica apoyo a aprendizaje y productividad junto con problemas de integridad, dependencia, fiabilidad y necesidad de alfabetización crítica. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: GPTutor favoreció engagement y apoyo personalizado; la calidad de interacción y diseño de prompts/tutoría condicionó la experiencia. La dependencia puede elevar una autoeficacia aparente pero asociarse negativamente con logro; el apoyo docente modifica la relación. La adaptación del feedback a contexto y desempeño mejoró pertinencia de la intervención y apoyo al aprendizaje ubicuo. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El análisis profundo del discurso permitió identificar estados de aprendizaje y apoyar retroalimentación adaptada al estudiante. Se reportaron mejoras en confianza y autonomía; la autorregulación fue clave para convertir microcontenidos y GenAI en aprendizaje significativo. El grupo experimental superó al control en programación y matemáticas; mayor interacción se relacionó con progreso y 80% valoró el feedback adaptativo. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: La dependencia de GenAI se relacionó con menor pensamiento independiente/innovador/crítico y menor autorreflexión; ansiedad y expectativas operaron como mecanismos. El grupo ChatGPT mostró menor retención a largo plazo que el aprendizaje tradicional (57.5% vs 68.5% en la prueba reportada), sugiriendo

riesgo de descarga cognitiva. Plantea rediseñar actividades hacia análisis, evaluación y creación, evitando delegar a GenAI los niveles cognitivos que se pretenden desarrollar. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: El uso estructurado de GenAI fortaleció evaluación crítica y uso responsable; también evidenció riesgos de dependencia y brechas de alfabetización en IA. ChatGPT mejoró calidad, elaboración, originalidad y autoeficacia y redujo esfuerzo mental; se observaron riesgos de calibración metacognitiva. La efectividad depende de segmentación, oportunidad, feedback y alineación con objetivos; propone marco para integrar microlearning en experiencias más amplias. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes. Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

En el subconjunto de evidencia correspondiente a este eje se observaron resultados que, considerados conjuntamente, pueden resumirse de la siguiente manera: Identifica que accesibilidad de evaluación digital exige compatibilidad con tecnologías de apoyo, autonomía e integridad bajo restricciones de examen. La literatura converge en alfabetización en IA, verificación de información y diseño de tareas que obliguen a justificar, contrastar y reflexionar sobre outputs de IA. Se obtuvo un modelo de 22 variables en cuatro dimensiones con propiedades psicométricas satisfactorias. La lectura comparada indica que estos hallazgos no deben interpretarse como equivalentes, porque proceden de diseños y métricas diferentes.

Tabla 2. Síntesis de hallazgos relevantes

Estudio	Hallazgo recuperado	Pertinencia
Exploring the Impact of Generative AI ChatGPT on Critical Thinking in Higher Education: Pa	El uso colaborativo y reflexivo favoreció fases de presencia cognitiva; el uso pasivo dirigido por IA mostró menor elaboración crítica.	Incluido por pertinencia directa con la
Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, featu	GPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados.	Incluido por pertinencia directa con la
Can Generative AI Revolutionise Academic Skills Development in Higher Education? A Systema	La mayoría reportó mejoras potenciales en habilidades cognitivas mediante feedback/ personalización, aunque persisten riesgos de dependencia y calidad	Incluido por pertinencia directa con la
Generating In-Context, Personalized Feedback for Intelligent Tutors with Large Language Mo	Diagnóstico cercano/superior al 80% en varios tipos de error, pero una fracción importante de pistas fue demasiado general, incorrecta o reveló la res	Incluido por pertinencia directa con la
Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect	La equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales.	Incluido por pertinencia directa con la
An AI-assisted critical thinking intervention to enhance undergraduate EFL learners' writi	La intervención reportó mejora en dimensiones de escritura vinculadas con pensamiento crítico y una aceptación favorable de la asistencia de IA.	Incluido por pertinencia directa con la
ChatGPT Interventions in Higher Education: A Systematic Review of Experimental Studies	Los efectos son positivos pero heterogéneos; intervenciones de duración media tienden a mostrar mejores resultados y los efectos sobre habilidades sup	Incluido por pertinencia directa con la
Systematic review of ChatGPT in higher education: Navigating impact on learning, wellbeing	El uso puede mejorar acceso y apoyo, pero la sobredependencia se asocia con riesgos para pensamiento crítico, autonomía y autenticidad académica.	Incluido por pertinencia directa con la
ChatGPT in higher education: A systematic literature review and research challenges	Identifica apoyo a aprendizaje y productividad junto con problemas de integridad, dependencia, fiabilidad y necesidad de alfabetización crítica.	Incluido por pertinencia directa con la
Promoting student engagement with GPTutor: An intelligent tutoring system powered by gener	GPTutor favoreció engagement y apoyo personalizado; la calidad de interacción y diseño de prompts/tutoría condicionó la experiencia.	Incluido por pertinencia directa con la

Nota. Síntesis textual basada exclusivamente en los resultados registrados en la matriz.

Sin embargo, su convergencia temática permite identificar condiciones recurrentes y formular una interpretación de conjunto centrada en la relación entre calidad del diseño, respuesta de los estudiantes y utilidad de la información producida para orientar decisiones educativas.

La síntesis muestra que la evidencia favorable se distribuye entre distintos tipos de resultados y que la pertinencia de cada estudio depende de su proximidad con el constructo central. Los hallazgos no se interpretan como intercambiables: algunos describen rendimiento, otros propiedades de medición, experiencia estudiantil, accesibilidad o desempeño de modelos. La comparación permite reconocer convergencias sin perder la especificidad de cada resultado y constituye la base para la discusión de mecanismos y condiciones de implementación.

Discusión

La síntesis permite sostener que el problema de efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios debe interpretarse desde una lógica de mecanismos de retroalimentación y no desde una oposición simplificada entre tecnología eficaz e ineficaz. Los estudios revisados muestran que el desempeño de una estrategia depende de la interacción entre propósito evaluativo, características de los estudiantes, diseño de la tarea, calidad de la medición y capacidad institucional para utilizar la información generada. Esta conclusión desplaza la atención desde la herramienta aislada hacia el sistema de evaluación en el que se inserta.

Una consecuencia importante es que la innovación tecnológica no elimina los problemas clásicos de validez, fiabilidad, equidad y utilidad. Por el contrario, puede hacerlos más visibles al ampliar la escala y velocidad de las decisiones. La evidencia revisada sugiere que las mejores implementaciones son aquellas que hacen explícitos los criterios, preservan oportunidades de retroalimentación y mantienen mecanismos de revisión humana cuando las consecuencias de una decisión son relevantes. En este sentido, la digitalización debe entenderse como una transformación del proceso evaluativo y no como una sustitución automática del juicio académico.

La heterogeneidad observada también explica por qué no resulta adecuado reducir el campo a un único tamaño de efecto. Las investigaciones difieren en constructos, duración, instrumentos, disciplinas, muestras y métricas. Esa variación no constituye únicamente una limitación; también aporta información sobre los contextos en los que una estrategia funciona y sobre las condiciones que deben ser consideradas antes de transferirla. La síntesis comparativa permite reconocer patrones de consistencia sin borrar diferencias que son esenciales para la toma de decisiones educativas.

Desde el punto de vista metodológico, el corpus evidencia una transición desde estudios centrados en adopción y percepción hacia diseños que buscan demostrar calidad de medición, impacto en resultados y sostenibilidad de la intervención. Sin embargo, persisten vacíos relacionados con comparaciones longitudinales, replicación entre instituciones, análisis de subgrupos y documentación de efectos no previstos. La agenda futura debería priorizar diseños que conecten resultados de aprendizaje con procesos de implementación y que informen no solo si una intervención funciona, sino para quién, bajo qué condiciones y con qué costos pedagógicos u organizacionales.

La principal contribución de esta revisión consiste en organizar la evidencia de forma que los resultados puedan utilizarse para decisiones académicas sin sobreextender sus alcances. La convergencia entre estudios ofrece argumentos para adoptar prácticas prometedoras, pero la diversidad metodológica exige preservar una lógica de evaluación continua. Las instituciones deberían tratar cada implementación como una intervención susceptible de seguimiento, con indicadores definidos antes de su despliegue y mecanismos para revisar resultados, sesgos, experiencia estudiantil y coherencia con los objetivos curriculares.

La investigación¹ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta el uso colaborativo y reflexivo favoreció fases de presencia cognitiva; el uso pasivo dirigido por IA mostró menor elaboración crítica. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

No² ofrece un punto de contraste útil para esta interpretación porque su estudio documenta gPT-4o ajustado alcanzó la mayor precisión y equidad general entre los modelos comparados. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como

evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

La evidencia sintetizada³ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la mayoría reportó mejoras potenciales en habilidades cognitivas mediante feedback/personalización, aunque persisten riesgos de dependencia y calidad de razonamiento.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

La contribución examinada⁴ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta diagnóstico cercano/superior al 80% en varios tipos de error, pero una fracción importante de pistas fue demasiado general, incorrecta o reveló la respuesta.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

No⁵ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la equidad varió por grupo demográfico y nivel de competencia; las medidas distinguen sesgo predictivo de diferencias reales.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

La investigación⁶ ofrece un punto de contraste útil para esta interpretación porque su estudio documenta la intervención reportó mejora en dimensiones de escritura vinculadas con pensamiento crítico y una aceptación favorable de la asistencia de IA.. Este hallazgo adquiere mayor significado cuando se compara con el resto del corpus: no actúa como evidencia autosuficiente, sino como una observación que ayuda a precisar mecanismos, límites y condiciones de aplicación. La lectura sistemática permite así evitar dos extremos frecuentes, la generalización excesiva a partir de un solo estudio y la pérdida de información contextual cuando se sintetizan resultados heterogéneos.

Implicaciones para la evaluación universitaria

Las implicaciones prácticas derivadas de la revisión deben traducirse en decisiones graduales. Para efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios, una implementación responsable requiere definir previamente el constructo que se desea medir, establecer indicadores de éxito, identificar grupos potencialmente afectados y prever mecanismos de revisión. La evidencia disponible favorece estrategias que integran la innovación con acompañamiento docente y evaluación continua de resultados.

En el plano institucional, la adopción debería acompañarse de protocolos de gobernanza de datos, criterios de transparencia y mecanismos de comunicación con estudiantes y docentes. Estas condiciones aumentan la interpretabilidad de los resultados y reducen el riesgo de utilizar indicadores fuera del contexto para el que fueron diseñados. La evaluación de una herramienta debe considerar tanto desempeño técnico como consecuencias pedagógicas y distributivas.

Para la investigación, el siguiente paso consiste en conectar mejor los estudios de eficacia con los de implementación. Resulta necesario saber no solo si una estrategia produce cambios, sino qué componentes explican esos cambios, qué recursos demanda y cómo se comporta en poblaciones diversas. Esa orientación permitiría acumular evidencia de forma más coherente y reducir la fragmentación que actualmente caracteriza parte de la literatura.

Limitaciones

La principal limitación del corpus es su heterogeneidad. Las diferencias entre diseños, instrumentos, poblaciones y desenlaces restringen la posibilidad de producir un estimador común y obligan a interpretar los patrones desde una lógica comparativa.

Esta limitación no invalida la síntesis, pero sí condiciona el alcance de las conclusiones y refuerza la necesidad de replicaciones con protocolos comparables.

También se observa una concentración de estudios en determinados contextos y tecnologías, lo que reduce la certeza sobre la transferibilidad a instituciones con infraestructura, perfiles estudiantiles o culturas de evaluación diferentes. Las investigaciones futuras deberían documentar con mayor precisión condiciones de implementación, características de los participantes y decisiones adoptadas durante el despliegue de las herramientas.

Finalmente, la revisión se apoya en el corpus consolidado documentado en las matrices suministradas. Las conclusiones representan fielmente ese conjunto y no se extienden a registros que no forman parte de la evidencia disponible. Esta delimitación preserva la trazabilidad y permite que futuras actualizaciones incorporen nuevos estudios sin alterar retrospectivamente la base analítica utilizada en el presente manuscrito.

Conclusiones

La revisión sistemática muestra que efectividad de la inteligencia artificial generativa en la evaluación formativa y retroalimentación del aprendizaje en estudiantes universitarios constituye un campo con evidencia suficiente para identificar patrones, pero todavía heterogéneo en diseños, poblaciones y métricas. El resultado central es que la calidad de las decisiones depende menos de la presencia aislada de una tecnología o instrumento que de su integración con objetivos de aprendizaje, criterios explícitos y procedimientos de seguimiento. La perspectiva de mecanismos de retroalimentación permite comprender por qué intervenciones similares pueden producir resultados diferentes y por qué la evaluación de efectividad debe considerar simultáneamente resultados, condiciones y calidad de la medición.

En términos aplicados, las instituciones de educación superior deberían adoptar una lógica gradual de implementación, con pruebas controladas, documentación de resultados y revisión periódica de efectos sobre distintos grupos de estudiantes. La evidencia respalda el uso de estrategias digitales cuando existe alineación pedagógica y cuando la información generada se utiliza para mejorar la enseñanza y el aprendizaje. También muestra la necesidad de evitar decisiones automáticas basadas en indicadores únicos, especialmente cuando las consecuencias académicas son relevantes o cuando existen diferencias de acceso, experiencia tecnológica o necesidades de apoyo.

La agenda de investigación debe avanzar hacia estudios longitudinales, comparaciones multicéntricas y diseños que documenten con mayor precisión mecanismos y efectos diferenciales. También se requiere mejorar la comparabilidad de indicadores y la transparencia en la descripción de instrumentos, algoritmos y procedimientos de intervención. Estas prioridades permitirán transformar un campo caracterizado por rápida innovación en una base de evidencia más acumulativa, replicable y útil para políticas universitarias. La revisión aporta una estructura para ese avance al separar hallazgos consistentes, condiciones de implementación y vacíos que todavía requieren investigación.

Referencias

1. Nesma Ragab Nasr; Chih-Hsiung Tu; Jennifer Werner; Tonia Bauer; et al. Exploring the Impact of Generative AI ChatGPT on Critical Thinking in Higher Education: Passive AI-Directed Use or Human-AI Supported Collaboration? *Education Sciences*. 2025;15(9):1198. <https://doi.org/10.3390/educsci15091198>
2. Accuracy and fairness of generative AI in automated essay scoring: Comparing GPT-4o, feature-based models, and human raters. *Assessing Writing*. 2026;69:101047. <https://doi.org/10.1016/j.asw.2026.101047>
3. Kangwa Daniel; M. M. Msambwa; Z. Wen. Can Generative AI Revolutionise Academic Skills Development in Higher Education? A Systematic Literature Review. *European Journal of Education*. 2025;60(1):e70036. <https://doi.org/10.1111/ejed.70036>
4. Jennifer M. Reddig; Arav Arora; Christopher J. MacLellan. Generating In-Context, Personalized Feedback for Intelligent Tutors with Large Language Models. *International Journal of Artificial Intelligence in Education*. 2025;35:3459–3500. <https://doi.org/10.1007/s40593-025-00505-6>
5. Assessing fairness in AI-assisted writing scoring: Developing fairness measures to detect predictive bias in automated essay scoring. *Assessing Writing*. 2026. <https://doi.org/10.1016/j.asw.2026.101066>
6. Xue Yin; Kun Dou; Jinlong Han. An AI-assisted critical thinking intervention to enhance undergraduate EFL learners' writing proficiency. *Studies in Educational Evaluation*. 2025;86:101480. <https://doi.org/10.1016/j.stueduc.2025.101480>
7. Yiwen Jin; Lies Sercu. ChatGPT Interventions in Higher Education: A Systematic Review of Experimental Studies. *Journal of Computer Assisted Learning*. 2025;41(4):e70072. <https://doi.org/10.1111/jcal.70072>
8. Systematic review of ChatGPT in higher education: Navigating impact on learning, wellbeing, and collaboration. *Social Sciences & Humanities Open*. 2025. <https://doi.org/10.1016/j.ssaho.2025.101866>
9. Maria Ijaz Baig; Elaheh Yadegaridehkordi. ChatGPT in the higher education: A systematic literature review and research challenges. *International Journal of Educational Research*. 2024;127:102411. <https://doi.org/10.1016/j.ijer.2024.102411>
10. Haoran Bai; Wing Cheung Lui; Paul Vinod Khatani. Promoting student engagement with GPTutor: An intelligent tutoring system powered by generative AI. *International Journal of Educational Technology in Higher Education*. 2025;22:77. <https://doi.org/10.1186/s41239-025-00571-9>
11. The paradox of self-efficacy and technological dependence: Unraveling generative AI's impact on university students' task completion. *The Internet and Higher Education*. 2025;65:100978. <https://doi.org/10.1016/j.iheduc.2024.100978>
12. Marta S. Tabares; Paola Vallejo; Alex Montoya; Daniel Correa. A feedback model applied in a ubiquitous microlearning environment using SECA rules. *Journal of Computing in Higher Education*. 2022;34:462–488. <https://doi.org/10.1007/s12528-021-09306-x>
13. Matt Grenander; Robert Belfer; Ekaterina Kochmar; Iulian V. Serban; François St-Hilaire; Jackie C. K. Cheung. Deep Discourse Analysis for Generating Personalized Feedback in Intelligent Tutor Systems. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021;35(17):15534–15544. <https://doi.org/10.1609/aaai.v35i17.17829>
14. Lucas Kohnke; Di Zou; Haoran Xie. Microlearning and generative AI for pre-service teacher education: a qualitative case study. *Education and Information Technologies*. 2025;30(15):21221–21248. <https://doi.org/10.1007/s10639-025-13606-5>
15. William Villegas-Ch; Diego Buenano-Fernandez; Alexandra Maldonado Navarro; Aracely Mera-Navarrete. Adaptive intelligent tutoring systems for STEM education: analysis of the learning impact and effectiveness of personalized feedback. *Smart Learning Environments*. 2025;12:41. <https://doi.org/10.1186/s40561-025-00389-y>
16. Ting Huang; Chenze Wu. The chain mediating effect of academic anxiety and performance expectations between academic self-efficacy and generative AI reliance. *Computers and Education Open*. 2025;9:100275. <https://doi.org/10.1016/j.caeo.2025.100275>
17. André Barcaui. ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*. 2025;12:102287. <https://doi.org/10.1016/j.ssaho.2025.102287>
18. Chahna Gonsalves. Generative AI's Impact on Critical Thinking: Revisiting Bloom's Taxonomy. *Journal of Marketing Education*. 2026;48(1). <https://doi.org/10.1177/02734753241305980>
19. Mahmoud Elkhodr; Ergun Gide. The SAGE framework for developing critical thinking and responsible generative AI use in cybersecurity education. *Discover Education*. 2025;4:517. <https://doi.org/10.1007/s44217-025-00935-3>
20. Marek Urban; Filip Děchtěrenko; Jiří Lukavský; Veronika Hrabalová; Filip Svacha; Cyril Brom; Kamila Urban. ChatGPT improves creative problem-solving performance in university students: An experimental study. *Computers & Education*. 2024;215:105031. <https://doi.org/10.1016/j.compedu.2024.105031>
21. Wali Khan Monib; Atika Qazi; Rosyzie Anna Apong. Microlearning beyond boundaries: A systematic review and a novel framework for improving learning outcomes. *Heliyon*. 2025;11(2):e41413. <https://doi.org/10.1016/j.heliyon.2024.e41413>
22. Hamida Al-Maamari; Manjit Singh Sidhu; Shaik Mazhar Hussain; Abbas M. Al-Ghaili; Kirandeep Kaur Sidhu. Accessible ICT-Enabled Examination Technologies for Students with Special Educational Needs in Higher Education: A Technical Evaluation of Usability, Reliability, Accommodation Quality, and Assessment Fairness. *International Journal of Special Education*. 2026;41(17s):635–654.
23. Achmad Salido; Irman Syarif; Melyani Sari Sitepu; Suparjan; Prima Rias Wana; Ryan Taufika; Rahyuni Melisa. Integrating critical thinking and artificial intelligence in higher education: A bibliometric and systematic review of skills and strategies. *Social Sciences & Humanities Open*. 2025;12:101924. <https://doi.org/10.1016/j.ssaho.2025.101924>

24. Agustín Mejías-Acosta; Mayra D'Armas Regnault; Eduardo Vargas-Cano; Jesennia Cárdenas-Cobo; Cristian Vidal-Silva. Assessment of digital competencies in higher education students: development and validation of a measurement scale. *Frontiers in Education*. 2024. <https://doi.org/10.3389/feduc.2024.1497376>

Financiación

Ninguna.

Conflictos de interés

No existen.

Contribución de autoría

Conceptualización: Luis David Bastidas González

Curación de datos Luis David Bastidas González

Análisis formal: Luis David Bastidas González

Investigación: Luis David Bastidas González

Metodología: Luis David Bastidas González

Supervisión: Luis David Bastidas González

Validación: Luis David Bastidas González

Visualización: Luis David Bastidas González

Redacción – borrador original: Luis David Bastidas González.

Redacción – revisión y edición: Luis David Bastidas González