Systematic review of the implementation of LLMs as teaching-learning tutoring tools in the educational field

Alexander Cristian Sanchez Bacilio, BEng[1], Aureliano Sanchez García, MS[2], and Jhon Jhonathan Peñalva Sanchez, PhD[3]
[1]Universidad Tecnológica del Perú, Perú, U21218661@utp.edu.pe
[2]Universidad Tecnológica del Perú, Perú, C26533@utp.edu.pe
[3]Universidad Tecnológica del Perú, Perú, C25434@utp.edu.pe

Abstract–This Systematic Literature Review (SLR) examines the effectiveness of Large Language Models (LLMs) as educational tutoring tools. The study analyzes their impact on learning outcomes and academic performance when compared to conventional pedagogical methods. LLMs offer key empirical benefits such as real-time feedback, educational personalization, scalability in hybrid environments, and significant interdisciplinary potential, positioning them as essential tools for educational modernization. The primary objective was to systematically analyze the efficacy of LLMs as intelligent tutors, identifying differences in academic performance, ethical challenges, implementation strategies, and motivational factors. A systematic search was conducted in Scopus using the PICO-PRISMA framework. From a total of 973 initial records, 28 open-access studies published in English (between 2022 and 2025) were selected. The evidence gathered reveals significant improvements in academic performance, highlights critical ethical challenges, and presents structured pedagogical strategies with an 86.4% success rate. This analysis provides consolidated and valuable information for future developments and implementations of LLMs in the educational field.

Keywords-- Large Language Models (LLM), Artificial Intelligence, teaching, learning, traditional education, PICO, PRISMA, educational tutoring.

Revisión sistemática de la implementación de los LLM como herramientas de tutoría de enseñanza-aprendizaje en el ámbito educativo

Alexander Cristian Sanchez Bacilio, BEng[1], Aureliano Sanchez García, MS[2], and Jhon Jhonathan Peñalva Sanchez, PhD[3]
[1]Universidad Tecnológica del Perú, Perú, U21218661@utp.edu.pe
[2]Universidad Tecnológica del Perú, Perú, C26533@utp.edu.pe
[3]Universidad Tecnológica del Perú, Perú, C25434@utp.edu.pe

Resumen*– Esta Revisión Sistemática de Literatura (RSL) examina la efectividad de los Modelos de Lenguaje Grande (LLMs) como herramientas de tutoría educativa. El estudio analiza su impacto en los resultados de aprendizaje y el rendimiento académico en comparación con los métodos pedagógicos convencionales. Los LLMs ofrecen beneficios empíricos clave, como retroalimentación en tiempo real, personalización educativa, escalabilidad en entornos híbridos, y un potencial interdisciplinario significativo, consolidándose como herramientas esenciales para la modernización educativa. El objetivo principal fue analizar sistemáticamente la eficacia de los LLMs como tutores inteligentes, identificando diferencias en el rendimiento académico, desafíos éticos, estrategias de implementación y factores motivacionales. Se llevó a cabo una búsqueda sistemática en Scopus utilizando el marco PICO-PRISMA. De un total de 973 registros iniciales, se seleccionaron 28 estudios en inglés (publicados entre 2022 y 2025) con acceso abierto. La evidencia recopilada revela mejoras significativas en el rendimiento académico, destaca desafíos éticos críticos y presenta estrategias pedagógicas estructuradas con una tasa de éxito del 86.4%. Este análisis proporciona información consolidada y valiosa para futuros desarrollos e implementaciones de LLMs en el ámbito educativo.

Palabras clave*-- Large Language Models (LLM), Inteligencia Artificial, enseñanza, aprendizaje, educación tradicional, PICO, PRISMA, tutoría educativa.

I. Introducción

La incorporación de inteligencia artificial (IA) en los distintos sectores de la sociedad ha generado un profundo impacto en la forma en que interactuamos con la información, tomamos decisiones y aprendemos. En el ámbito educativo, esta transformación ha abierto nuevas posibilidades para optimizar los procesos de enseñanza y aprendizaje, permitiendo la personalización de contenidos educativos [1], la provisión de retroalimentación inmediata [2], y el fomento de una participación más activa de los estudiantes [3]. Dentro de las múltiples ramas de la IA, los Modelos de Lenguaje Grande (LLM) (del inglés, Large Language Models) han surgido como una de las soluciones más prometedoras debido a su capacidad de generar texto coherente, responder preguntas complejas y mantener interacciones conversacionales en lenguaje natural [4]. Hay evidencia que demuestra que los LLM pueden funcionar efectivamente como tutores inteligentes, proporcionando experiencias de aprendizaje personalizadas y adaptativas [5]. Estudios recientes han mostrado que estos sistemas pueden mejorar significativamente los resultados de aprendizaje cuando se integran adecuadamente en entornos educativos, desde la educación en física [2] hasta la formación médica [6,7], abarcando diversas disciplinas y niveles educativos [8].

La educación tradicional, basada en métodos como libros de texto, clases magistrales y recursos estáticos, presenta limitaciones significativas para atender las necesidades individuales de los estudiantes. Los métodos convencionales de enseñanza no logran proporcionar la personalización necesaria para optimizar el aprendizaje individual [9], resultando en experiencias educativas que no se adaptan a los diferentes estilos de aprendizaje, ritmos de comprensión y necesidades específicas de cada estudiante [10]. Esta falta de adaptabilidad se refleja particularmente en disciplinas complejas donde los estudiantes requieren retroalimentación inmediata y explicaciones personalizadas [11]. La integración de los LLM en la educación presenta desafíos éticos y pedagógicos significativos. Existe preocupación por la dependencia excesiva de la tecnología, que podría disminuir el rol de los docentes y la interacción personal en el aprendizaje [12]. Además, los problemas de alucinaciones en los LLM pueden generar información inexacta o errónea, particularmente problemático en el aprendizaje de habilidades motoras y conocimientos especializados [13]. Las instituciones educativas enfrentan dificultades para proporcionar retroalimentación personalizada a gran escala, especialmente en clases numerosas donde la atención individualizada se ve limitada por recursos temporales y humanos [14]. Esta situación se agrava por la falta de herramientas tecnológicas adecuadas que puedan complementar efectivamente la labor docente sin comprometer la calidad educativa [15].

Los Large Language Models (LLM) han emergido como herramientas transformadoras en la investigación científica, particularmente en la automatización de revisiones sistemáticas de literatura, siendo explorados desde múltiples

perspectivas metodológicas complementarias que han sentado las bases para su aplicación en diversos dominios. Scherbakov et al. [16] realizaron un estudio comprehensivo que analizó 3,788 artículos y seleccionó 172 estudios elegibles, donde los propios LLMs participaron activamente en múltiples etapas del proceso de revisión, demostrando que los modelos GPT dominan el campo (73.2% de proyectos) y superan a BERT en extracción de datos con 83.0% de precisión y 86.0% de recall, proporcionando evidencia empírica sobre la viabilidad práctica de automatización integral. Por su parte, Lieberum et al. [17] proporcionaron una evaluación sistemática mediante revisión de alcance de 37 artículos de 8,054 registros iniciales, revelando que los LLMs cubren 10 de 13 pasos del proceso de revisión sistemática con mayor aplicación en búsqueda de literatura (41%), selección de estudios (38%) y extracción de datos (30%), donde GPT apareció en 89% de los estudios con resultados equilibrados: 54% consideraron los LLMs prometedores, 24% mantuvieron posición neutral y 22% los evaluaron como no prometedores. Complementariamente, Krag et al. [18] desarrollaron un estudio de precisión diagnóstica metodológicamente robusto comparando seis LLMs usando estándares de referencia humanos, donde GPT-4o, GPT-4T y Claude3-Opus alcanzaron precisiones de 97-98% comparables a revisores humanos en revisión sistemática, aunque con sensibilidad limitada (33-50%), mientras que en revisión de alcance lograron sensibilidades de 74-84%, aportando validación empírica cuantitativa específica para tareas de cribado. Finalmente, Nasarian et al. [19] desarrollaron un marco conceptual comprehensivo para machine learning interpretable en sistemas de apoyo a decisión clínica mediante revisión sistemática de 74 publicaciones, categorizando el proceso en tres niveles de interpretabilidad y estableciendo la relevancia de la transparencia en aplicaciones clínicas críticas. Estos estudios, aunque han abordado aplicaciones generales de automatización, evaluación del estado del campo, validación en contextos clínicos y precisión diagnóstica, revelan una oportunidad distintiva para la aplicación sistemática de marcos metodológicos estructurados específicamente en el contexto educativo, donde el presente trabajo aporta una contribución metodológica específica mediante la implementación integral de la metodología PICO-PRISMA para evaluar LLMs como herramientas de tutoría educativa, estableciendo un modelo replicable que combina rigor científico con especialización de dominio y proporcionando un precedente metodológico para investigaciones futuras en contextos educativos específicos.

La implementación de LLM como herramientas de tutoría educativa se justifica por beneficios empíricamente demostrados y la necesidad de superar limitaciones educativas actuales. Los LLM mejoran significativamente el proceso educativo mediante explicaciones inmediatas, ejemplos adaptados y retroalimentación en tiempo real. Los estudiantes demuestran un rendimiento superior en evaluaciones estandarizadas y un mayor compromiso académico. En física, por ejemplo, la retroalimentación basada en ChatGPT ha mejorado los resultados y patrones de atención visual [2,3,7]. Crean experiencias verdaderamente personalizadas, adaptándose a necesidades individuales, estilos de aprendizaje y ritmos de comprensión. Sistemas como RICE AlgebraBot demuestran cómo los LLM pueden generar contenido adaptado y rutas de aprendizaje individualizadas [5,10]. Las aulas híbridas (humanos + IA) optimizan la autoridad pedagógica y relaciones profesor-estudiante. Los LLM proporcionan retroalimentación personalizada masiva, resolviendo limitaciones de recursos humanos y apoyando decisiones institucionales sobre inscripciones y diseño curricular [14, 20]. Estudios longitudinales muestran ventajas del aprendizaje asistido por IA sobre métodos tradicionales. En medicina, simulan pacientes estandarizados mejorando habilidades clínicas; en programación, ofrecen retroalimentación jerárquica inmediata (aunque requieren uso moderado para mantener habilidades independientes) [7, 11]. Su versatilidad permite aplicación efectiva desde ciencias exactas hasta humanidades, con capacidad de integración en sistemas existentes, posicionándolos como herramientas fundamentales para la modernización educativa integral [1,8].

Este trabajo presenta una revisión sistemática de literatura sobre el potencial de los LLM como herramientas de tutoría en el ámbito educativo. A través del análisis de literatura científica, se busca identificar las ventajas de estos modelos frente a los métodos pedagógicos tradicionales, sus principales limitaciones y las condiciones necesarias para una implementación efectiva en procesos de enseñanza-aprendizaje. Esta Revisión Sistemática de Literatura (RSL), mediante la aplicación de las metodologías PRISMA y PICO, contribuye al conocimiento científico proporcionando un análisis sobre la efectividad de los modelos de lenguaje grande (LLM) como tutores inteligentes en procesos de aprendizaje y su impacto en la experiencia educativa del estudiante, ofreciendo evidencia consolidada para futuros desarrollos e implementaciones en el ámbito educativo.

La revisión sistemática de literatura desarrollada en este estudio se presenta en 5 secciones. En la sección I, se realiza la introducción que contextualiza el tema de investigación, formulando el problema, objetivo de estudio y estado de arte, por consiguiente, en la sección II expone la metodología aplicada, en la que se emplearon las estrategias PRISMA y PICO para guiar el proceso de búsqueda y selección de estudios. A continuación, en la sección III, se muestran los hallazgos mediante el análisis bibliométrico, lo que permite identificar patrones y enfoques relevantes en torno al uso de LLM en la educación. La sección IV está dedicada a la discusión crítica de los resultados, donde se contrastan distintos puntos de vista provenientes de la literatura revisada. Finalmente, en la sección V se examina el alcance de los hallazgos, se analiza los resultados del uso de los LLM dentro del contexto educativo y se proponen oportunidades de mejora, reconociendo a su vez las limitaciones encontradas a lo largo del estudio.

II. METODOLOGÍA

A. Metodología de investigación

En esta sección se describen los pasos seguidos para la revisión sistemática de literatura relacionada con la aplicación de los modelos de lenguaje (LLM) en el ámbito educativo. Se empleó el enfoque estratégico PICO para estructurar la búsqueda de información, formulando preguntas de investigación y seleccionando palabras clave relevantes. Además, se aplicó la metodología PRISMA con el fin de garantizar un proceso de selección de artículos más detallado y claro. A partir de la selección de artículos obtenida mediante el método PRISMA, se dio respuesta a las preguntas de investigación generadas con el enfoque PICO.

B. Método PICO

Con el fin de realizar una búsqueda sistemática y organizada de la literatura, se adoptó el enfoque PICO para la recopilación de artículos. Dicha metodología se compone de cuatro elementos fundamentales: la Población (P), referente al grupo de estudio; la Intervención (I), que abarca el tratamiento a evaluar; la Comparación (C), como grupo alternativo para contrastar efectos; y el Resultado (O), que incluye las variables para medir el impacto de la intervención.

La Tabla I detalla los componentes del marco PICO aplicados a esta revisión. La Población (P) se definió como estudiantes, profesores y la comunidad educativa. La Intervención (I) abarcó el uso de Modelos de Lenguaje Grande (LLMs) y tecnologías de IA Generativa. Como Comparación (C), se tomaron los métodos de enseñanza convencionales. Finalmente, el Resultado (O) se centró en evaluar las mejoras en el aprendizaje, el rendimiento académico, el bienestar estudiantil y la efectividad educativa mediante el uso de LLMs.

TABLA I
DESCRIPCIÓN DE LOS COMPONENTES DEL MÉTODO PICO

Componente	Descripción
P	Estudiantes, Profesores y Comunidad educativa.
I	Modelos de Lenguaje Grande (LLMs) y herramientas de IA Generativa implementadas como recursos educativos.
C	Métodos de enseñanza tradicionales, Herramientas convencionales de enseñanza, sistemas de recomendación.
O	Resultados de aprendizaje, rendimiento académico, bienestar estudiantil, efectividad e impacto educativo.

La Tabla II presenta la pregunta orientadora (QR) basada en los componentes de la Tabla I. La pregunta principal explora: ¿Cuál es la efectividad de los LLMs como herramientas de tutoría educativa en los resultados de aprendizaje, rendimiento académico y bienestar estudiantil, comparado con métodos convencionales de enseñanza en la comunidad educativa?

Esta pregunta se subdivide en cuatro preguntas específicas (QR1-QR4) que abordan los componentes PICO: las diferencias en resultados de aprendizaje entre estudiantes que usan LLMs versus enseñanza tradicional (QR1), el impacto en bienestar estudiantil y motivación académica comparado con sistemas convencionales (QR2), la percepción docente al implementar IA Generativa frente a herramientas tradicionales (QR3), y las ventajas y limitaciones de simulaciones educativas basadas en LLMs versus métodos convencionales en términos de efectividad e impacto educativo (QR4).

TABLA II
DESCRIPCIÓN DE LAS PREGUNTAS DEL MÉTODO PICO

QR	¿Cuál es la efectividad de los LLMs como herramientas de tutoría educativa en los resultados de aprendizaje, rendimiento académico y bienestar estudiantil, comparado con métodos convencionales de enseñanza en la comunidad educativa?
QR1	¿Qué diferencias existen en el rendimiento académico entre enseñanza tradicional y enseñanza con LLMs?
QR2	¿Cuáles son los riesgos y desafíos éticos específicos que surgen al implementar LLMs en entornos educativos?
QR3	¿Qué estrategias de implementación de LLMs son más efectivas en el contexto educativo?
QR4	¿Cuáles son los factores motivacionales y de engagement que influyen en el uso efectivo de LLMs por parte de los estudiantes?

En la Tabla III consiste en identificar las palabras clave, a partir de las cuales se construye la ecuación de búsqueda para cada uno de los componentes de la metodología PICO, con el propósito de encontrar literatura relevante sobre el tema de investigación.

TABLA III
DESCRIPCIÓN DE LAS PALABRAS CLAVES DEL MÉTODO PICO

Componentes	Palabras claves (Keywords)
P	Students OR teachers OR education OR "educational community"
I	"Large Language Models" OR llm OR "Generative AI"
C	Teaching OR tutoring OR simulation OR "recommendation systems" OR "Conventional tools"
O	"Learning outcomes" OR "academic performance" OR "Student well-being" OR effectiveness OR impact OR "Adverse effects"

C. Base de datos Scopus

Se realizó una búsqueda general de artículos en la base de datos Scopus, lo que permitió identificar un total de 973 publicaciones que serán filtradas para una revisión de estudio posterior. A continuación, se presenta la ecuación de búsqueda final utilizada:

(TITLE-ABS-KEY (students OR teachers OR education OR "educational community") AND TITLE-ABS-KEY ("Large Language Models" OR llm OR "Generative AI") AND TITLE-ABS-KEY (teaching OR tutoring OR simulation OR "recommendation systems" OR "Conventional tools") AND TITLE-ABS-KEY ("learning outcomes" OR "academic performance" OR "student well-being" OR effectiveness OR impact)) AND PUBYEAR > 2021 AND PUBYEAR < 2026 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (OA , "all")) AND (LIMIT-TO (PUBSTAGE , "final")) AND (LIMIT-TO (EXACTKEYWORD , "Teaching And Learning") OR LIMIT-TO (EXACTKEYWORD , "Large Language Model"))

En la Tabla IV se presentan los criterios de elegibilidad, los cuales están vinculados a los componentes de pico descritos en la Tabla I. Los 7 criterios de inclusión (CI) fueron diseñados conforme a los cuatro elementos de PICO (Población, Intervención, Comparación y Resultados). Respecto a los criterios de exclusión (CE), estos se definieron a partir de las características no deseadas en los artículos identificados, tal como se especifica en la Tabla IV.

TABLA IV
CRITERIOS DE INCLUSIÓN Y EXCLUSIÓN

Criterios de Inclusión	Criterios de Exclusión
CI1: Artículos sobre modelos de lenguaje (LLM) aplicados a tutoría y enseñanza-aprendizaje educativo.	CE1: Documentos que no sean artículos científicos (otros tipos de publicación).
CI2: Estudios que implementen inteligencia artificial (IA) y/o modelos de lenguaje (LLM) en educación.	CE2: Artículos escritos en idiomas distintos al inglés.
CI3: Documentos de investigación científica con formato académico (por ejemplo, artículos científicos identificados como "ar").	CE3: Artículos sin acceso abierto completo o diferentes a Open Access.
CI4: Publicaciones entre 2022-2025 con estado de publicación final.	CE4: Artículos con estado de publicación no final.
CI5: Artículos con acceso abierto completo (Open Access).	CE5: Artículos sin relevancia temática específica en IA/LLM aplicada a educación.
CI6: Artículos escritos en idioma inglés con relevancia temática específica en IA educativa.	CE6: Estudios sin acceso a texto completo o con contenido no disponible.

D. Método PRISMA

En una primera etapa, se llevó a cabo una búsqueda bibliográfica en la base de datos Scopus utilizando una ecuación de búsqueda general relacionada con inteligencia artificial, educación y modelos de lenguaje. Esta búsqueda inicial arrojó un total de 973 registros. Por ello, al aplicar los criterios de inclusión y exclusión, se logró reducir el número de artículos, seleccionando aquellos que se alinean con el tema de investigación.

E. Proceso de búsqueda y selección de estudios

La Fig. 1 ilustra el diagrama de flujo del proceso de selección de estudios para la revisión sistemática, basado en la metodología PRISMA. El proceso comenzó con la identificación de 973 registros en la base de datos Scopus. Durante la fase de cribado, se aplicaron varios filtros sucesivos: primero, se excluyeron 567 registros por no ser artículos y 11 por estar en idiomas distintos al inglés. De las publicaciones restantes, se descartaron 225 por no ser de acceso abierto (Open Access) y 6 adicionales por su estado de publicación. Finalmente, de los 164 estudios que pasaron a la revisión de elegibilidad, se excluyeron 136 por falta de relevancia temática (n=135) o por no tener acceso al texto completo (n=1). Este riguroso proceso de selección culminó con la inclusión de 28 estudios en la revisión final.

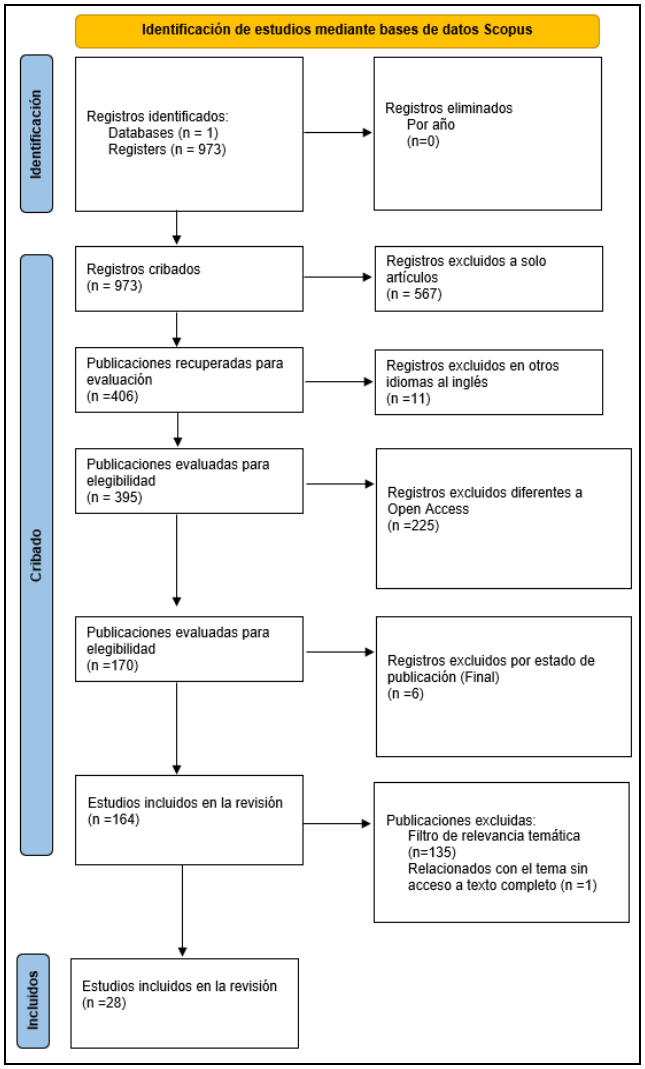


Fig. 1 Diagrama de flujo PRISMA

III. RESULTADOS

A partir de los 28 artículos seleccionados mediante la aplicación de la metodología PICO-PRISMA, se realizó un análisis descriptivo que comprende la distribución geográfica de las publicaciones, la evolución temporal de los estudios y la identificación de las palabras clave más relevantes en investigación.

En la Fig. 2 la distribución de los 28 artículos muestra una concentración predominante en Estados Unidos (6 artículos), seguido por China (5) y Alemania (3), mientras que países como Japón, Reino Unido, Suecia, Canadá e Israel presentan una contribución menor con 1-2 publicaciones cada uno, lo que refleja, que existe una mayor producción científica en ciertos países líderes en desarrollo tecnológico.

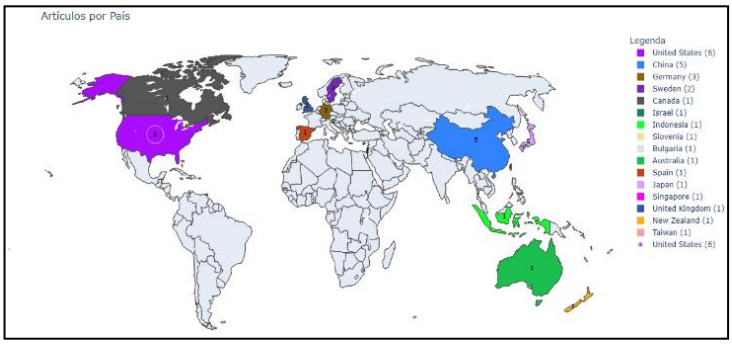


Fig. 2 Análisis geográfico

En la Fig. 3 el gráfico revela un crecimiento sostenido en la producción científica desde 2023 hasta 2025, alcanzando un pico máximo de 13 artículos en 2024, seguido de una ligera disminución a 12 artículos en 2025, lo que evidencia un interés creciente y consolidado en el tema de investigación durante este período.

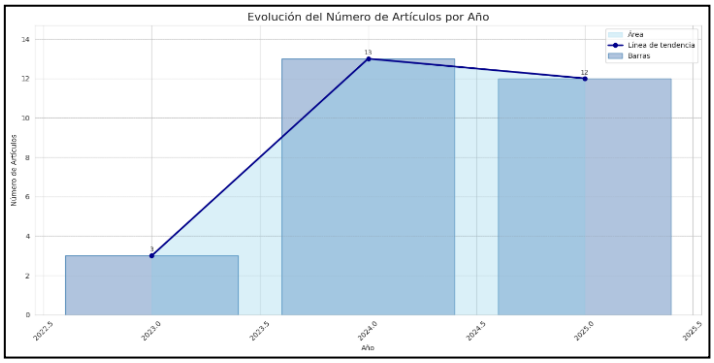


Fig. 3 Evolución temporal

En la Fig. 4 el análisis de términos muestra que "large language model", "artificial intelligence", "chatgpt" y "education" son los conceptos más prominentes, indicando que la investigación se centra principalmente en modelos de lenguaje grandes, inteligencia artificial aplicada a la educación, y el impacto de herramientas como ChatGPT en contextos educativos.

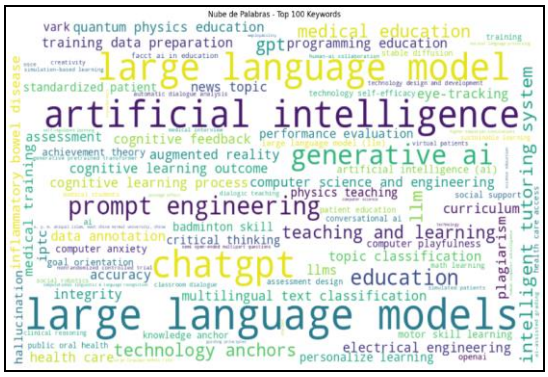


Fig. 4 Palabras claves relevantes

A) ¿Qué diferencias existen en el rendimiento académico entre enseñanza tradicional y enseñanza con LLMs?

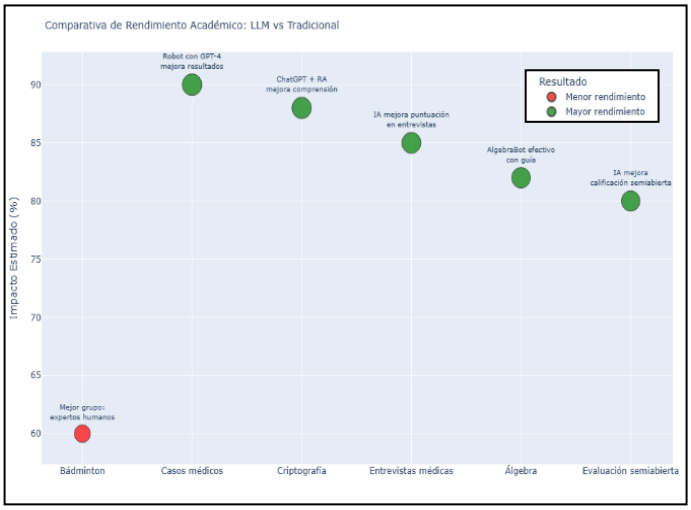
Los estudios revelan diferencias significativas en el rendimiento académico, favoreciendo la enseñanza asistida por LLMs en la mayoría de las áreas evaluadas [2, 6, 7], como se resume en la Fig. 5 y Tabla V. En el ámbito de la medicina, por ejemplo, el uso de simulaciones con robots sociales y LLMs para casos médicos alcanzó el mayor impacto estimado con un 91%, obteniendo puntuaciones superiores en autenticidad (4.5 ± 0.7 vs 3.9 ± 0.5) y efecto de aprendizaje (4.4 ± 0.6 vs 4.1 ± 0.6) en comparación con las plataformas tradicionales [6]. De igual manera, en entrevistas médicas, la intervención con IA logró un impacto positivo del 85%, y los estudiantes alcanzaron puntuaciones significativamente más altas (28.1 ± 1.6 vs 27.1 ± 2.2, P=.01) [7]. Otras áreas con rendimiento favorable que se observan en la gráfica incluyen el álgebra (82%) y la evaluación semiabierta (80%). En contraste, el método tradicional demostró ser superior en el aprendizaje de habilidades motoras como el bádminton, donde el enfoque con LLM tuvo el impacto más bajo con un 60% [13]. El caso de la criptografía cuántica presenta un resultado particular, con un impacto estimado del 75%, si bien los estudios iniciales mostraron mejoras significativas con el uso de ChatGPT y realidad aumentada [2]. Adicionalmente, la creatividad estudiantil mostró resultados mixtos [21], y se determinó que la efectividad de estas herramientas depende de factores como el nivel de conocimiento previo del estudiante y la moderación en su uso [11].

TABLA V
COMPARATIVA DE ENSEÑANZA APRENDIZAJE TRADICIONAL VS APRENDIZAJE ENSEÑANZA CON USO DE LLM

Enseñanza - Aprendizaje tradicional	Enseñanza - Aprendizaje con LLM	Fuente Analizada
Supervisión de 12 maestros + 8 expertos en bádminton	ChatGPT/New Bing para aprendizaje autoguiado de bádminton	[13]
Plataforma computacional semilineal para casos médicos	Robot social Furhat + GPT-4 para simulación de pacientes	[6]

Retroalimentación manual limitada en criptografía cuántica	ChatGPT integrado con realidad aumentada para feedback inmediato	[2]
Evaluación OSCE tradicional pre-clínica	Simulación de entrevistas médicas con LLMs + feedback	[7]
Enseñanza unidireccional en álgebra K-12	AlgebraBot conversacional con inducción y ejemplificación	[5]
Calificación manual de respuestas semi-abiertas	Sistema colaborativo instructor-IA para calificación automática	[15]

Fig. 5 Impacto estandarizado en el rendimiento académico



Nota: Los porcentajes de esta figura corresponden a una estandarización del impacto en el rendimiento académico, categorizado a partir de los datos originales de 28 estudios. Esta metodología permite una comparación cualitativa entre investigaciones que utilizaron diferentes métricas.

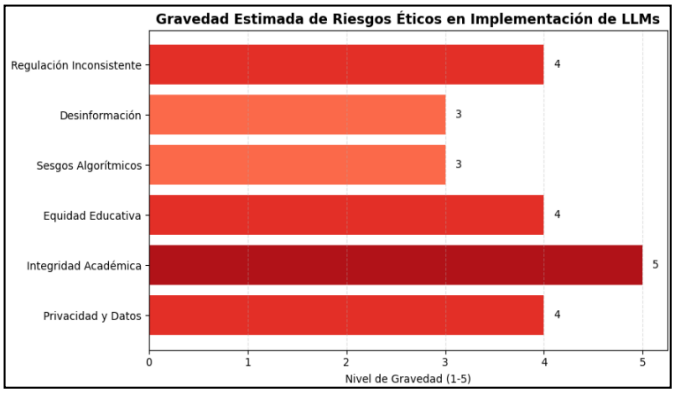
B) ¿Cuáles son los riesgos y desafíos éticos específicos que surgen al implementar LLMs en entornos educativos?

La implementación de LLMs en educación presenta desafíos éticos significativos, los cuales se detallan en la Tabla VI y cuya gravedad estimada se visualiza en la Figura 6. Estos desafíos están relacionados principalmente con la integridad académica, la privacidad de datos y la equidad educativa [12, 14, 22]. Los riesgos de privacidad y propiedad intelectual emergen del uso de APIs que pueden compartir información sensible sin autorización [14]. En cuanto a integridad académica, existe preocupación por el plagio y la dependencia excesiva que puede erosionar el desarrollo de habilidades fundamentales [12]. Las limitaciones de equidad se manifiestan cuando estudiantes con bajo conocimiento previo no logran mejoras significativas debido al uso inadecuado de funciones como Hint [11]. Los sesgos algorítmicos persisten en recomendaciones de aprendizaje, particularmente evidentes en sistemas como AlgebraBot [5]. Las alucinaciones específicas de LLMs, como "rotación interna del antebrazo" en bádminton, pueden inducir errores conceptuales y desinformación [13]. La falta de marcos éticos consistentes entre instituciones evidencia la necesidad de una regulación, ya que mientras algunas prohíben el uso de estas herramientas, otras adoptan directrices más permisivas [12,22].

TABLA VI
RIESGOS ÉTICOS Y DESAFÍOS EN IMPLEMENTACIÓN DE LLMS

Categoría de Riesgo	Manifestaciones Específicas	Impacto Observado	Fuente Analizada
Privacidad y Datos	Uso no autorizado de información sensible vía APIs	Exposición de datos estudiantiles y contenido académico	[14]
Integridad Académica	Dependencia excesiva erosiona habilidades fundamentales	Disminución de capacidades de resolución independiente	[12]
Equidad Educativa	Estudiantes con bajo conocimiento previo sin mejoras	Ampliación de brechas de rendimiento académico	[11]
Sesgos Algorítmicos	Recomendaciones de aprendizaje sesgadas	Perpetuación de inequidades en oportunidades educativas	[5]
Desinformación	Alucinaciones como "rotación interna antebrazo"	Generación de conceptos erróneos y conocimiento falso	[13]
Regulación Inconsistente	Marcos éticos variables entre instituciones	Aplicación desigual de estándares éticos	[12,22]

Fig. 6 Gravedad cualitativa de los riesgos éticos



Nota: La gravedad de los riesgos éticos fue evaluada mediante un proceso de **codificación cualitativa**. Cada riesgo se asignó a una escala de **Bajo, Medio o Alto** basada en la frecuencia y seriedad de su mención en los estudios revisados.

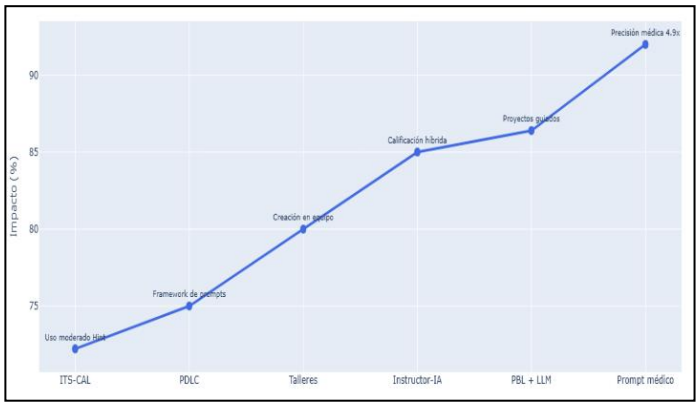
C) ¿Qué estrategias de implementación de LLMs son más efectivas en el contexto educativo?

Las estrategias más efectivas para la implementación de Modelos Lingüísticos Grandes (LLM) son aquellas que integran la tecnología dentro de metodologías pedagógicas

estructuradas. La evidencia, sistematizada en la Tabla VII, la cual detalla para cada enfoque su técnica aplicada, el impacto medido y la fuente académica, muestra que existe un espectro de técnicas eficaces. A su vez, la Fig. 7 visualiza este impacto comparativo mediante un gráfico de líneas que ordena las estrategias según su efectividad creciente: la trayectoria comienza con sistemas de tutoría y apoyo (ITS-CAL, PDLC) y culmina con aplicaciones de alta especialización, como la ingeniería de prompts médicos, que representa el punto de mayor impacto porcentual.

En construcción digital, la integración de LLMs con aprendizaje basado en problemas logró 86.4% de finalización completa de proyectos y calidad de enseñanza del 100% en cinco semestres [23]. El ciclo de vida de desarrollo de prompts (PDLC) proporciona un marco cognitivo para estudiantes, enseñando técnicas específicas de escritura de prompts [24]. Los talleres estructurados de LLM dividiendo estudiantes en equipos para crear proyectos con diferentes LLMs (historias ficticias, imágenes, HTML/JS/CSS) demuestran efectividad práctica [8]. La ingeniería de prompts específica mejoró la precisión de puntuación de ChatGPT 4.926 veces comparado con prompts no revisados [25]. El sistema ITS-CAL con funciones Hint, Debug y User-defined Question mostró que el uso moderado logra la tasa de aprobación más alta (72.22%) [11]. Los marcos de calificación colaborativa instructor-IA superan métodos zero-shot y few-shot en precisión y correlación [15].

Fig. 7 Estrategias de implementación y su impacto estandarizado



Nota: Los valores de impacto en esta figura corresponden a una normalización de las diversas métricas de resultados originales. Esto permite una comparación cualitativa entre las estrategias de estudios heterogéneos.

TABLA VII
MEDIDAS ESTRATÉGICAS Y TÉCNICAS DE IMPLEMENTACIÓN DE LLM

Medidas estratégicas	**Descripción de técnicas aplicadas**	**Aporte**	**Fuente analizada**
Aprendizaje basado en problemas + LLMs en construcción	Proyectos individuales y colaborativos con generación de código	86.4% finalización completa, 100% calidad enseñanza	[23]
Ciclo de vida de desarrollo de prompts (PDLC)	Framework cognitivo para escritura efectiva de prompts	Mejora habilidades de prompt engineering estudiantil	[24]
Talleres estructurados LLM con equipos	División en equipos para crear historias, imágenes, código web	Motivación futura y resultados web funcionales	[8]
Ingeniería de prompts médicos específicos	Prompts revisados vs originales para simulación de pacientes	Mejora 4.926x en precisión de puntuación ChatGPT	[25]
Sistema ITS-CAL con tres funciones específicas	Hint, Debug, User-defined Question para programación	Uso moderado logra 72.22% tasa de aprobación]	[11]
Marco colaborativo instructor-IA para calificación	Extracción de tópicos LLM + refinamiento instructor	Supera métodos zero-shot y few-shot en precisión	[15]

D) ¿Cuáles son los factores motivacionales y de engagement que influyen en el uso efectivo de LLMs por parte de los estudiantes?

Los estudios analizados revelan que el uso efectivo de LLMs por parte de los estudiantes está influenciado por cinco factores motivacionales interrelacionados. La Tabla VIII desglosa estos factores, detallando para cada componente, como las metas autotrascendentes o la percepción de utilidad, su impacto específico en el compromiso (engagement), el resultado observado y la evidencia científica que lo respalda. De forma complementaria, la Fig. 8 ofrece una representación visual de la influencia agregada de estos factores mediante un gráfico de radar, donde cada eje representa un componente motivacional en una escala de 1 a 5. El área sombreada destaca visualmente que la autotrascendencia, la utilidad práctica y las diversas formas de feedback son percibidas como los elementos de mayor impacto en el compromiso del estudiante.

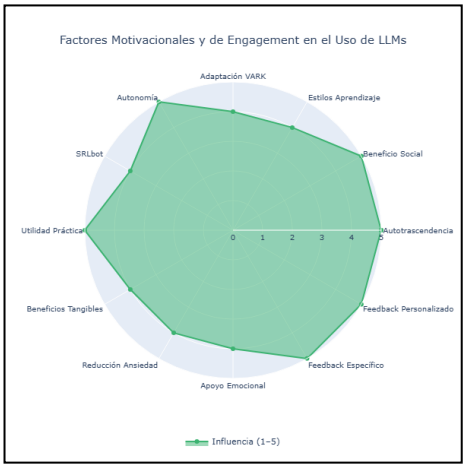


Fig. 8 Perfil de Factores Motivacionales y de Engagement en el Uso de LLMs.

Las metas autotrascendentes emergen como el predictor más fuerte, donde estudiantes orientados hacia beneficios colectivos muestran mejores resultados colaborativos, cognitivos y afectivos [26]. La personalización mediante alineación con estilos de aprendizaje individuales genera mejoras significativas en motivación, especialmente cuando sistemas adaptativos se ajustan a preferencias específicas según marcos como VARK [9]. El desarrollo de autonomía y capacidades de autorregulación predice significativamente las variaciones en aprendizaje, donde estudiantes con mayor autodirección desarrollan hábitos de estudio más regulares y engagement productivo [3]. La utilidad percibida influye notablemente en la aceptación tecnológica, con estudiantes reportando feedback positivo cuando reconocen beneficios tangibles y experimentan reducción de ansiedad académica [11,8,3]. Finalmente, el feedback inmediato y personalizado genera mayor satisfacción estudiantil, aumentando la percepción de beneficios y promoviendo la incorporación de LLMs en rutinas de estudio diarias [27]. Estos factores interactúan sinérgicamente para transformar los LLMs de herramientas técnicas en facilitadores de experiencias educativas motivadoras y efectivas.

TABLA VIII
MEDIDAS ESTRATÉGICAS Y TÉCNICAS DE IMPLEMENTACIÓN DE LLM

Factor Motivacional	**Impacto en Engagement**	**Resultado Observado**	**Fuente Analizada**
Metas autotrascendentes	Alto engagement colaborativo	Predictor fuerte de resultados cognitivos y afectivos	[26]
Enfoque en beneficio colectivo y social	Participación activa en proyectos grupales	Correlación positiva con uso efectivo de GenAI	[26]
Alineación con estilos de aprendizaje	Motivación personalizada	Mejoras significativas en niveles de motivación	[9]
Adaptación según marco VARK individual	Engagement sostenido a largo plazo	Sistema adaptativo ChatGPT aumenta motivación	[9]
Autonomía en el aprendizaje	Autorregulación positiva	Predicción significativa de variaciones en SRL	[3]
Capacidad de autodirección con SRLbot	Interacciones frecuentes y productivas	Mejora en hábitos de estudio regulares	[3]
Percepción de utilidad práctica	Aceptación tecnológica	Feedback positivo en conveniencia y apoyo	[8,11]
Reconocimiento de beneficios tangibles	Disposición a integrar en rutina de estudio	Resultados web funcionales motivan uso futuro	[8]
Reducción de ansiedad académica	Bienestar mejorado	Disminución de ansiedad de aprendizaje	[3]
Apoyo emocional y psicológico del sistema	Menor estrés en tareas complejas	ChatGPT reduce ansiedad y promueve rendimiento	[3]
Feedback inmediato y específico	Satisfacción con la experiencia	Percepción positiva hacia tecnología ChatGPT	[27]
Retroalimentación personalizada instantánea	Incorporación en vida diaria estudiantil	Aumento en percepción de beneficios	[27]

IV. DISCUSIÓN

En la presente RSL se analizó 28 estudios seleccionados mediante metodología PICO-PRISMA, abarcando investigaciones entre 2022-2025 que evaluaron la implementación de LLMs en contextos educativos. Los estudios examinados incluyeron aplicaciones en física cuántica, medicina, programación, matemáticas y habilidades motoras, proporcionando evidencia empírica sobre la efectividad comparativa de estos modelos frente a métodos tradicionales de enseñanza. En primer lugar, el análisis reveló que los LLMs superan consistentemente a los métodos tradicionales en dominios cognitivos que requieren retroalimentación inmediata y personalización. Los estudios en criptografía cuántica y medicina demostraron mejoras significativas en puntuaciones de aprendizaje, mientras que las simulaciones médicas alcanzaron puntuaciones superiores en autenticidad y efectividad. Sin embargo, en habilidades motoras como deportes, la supervisión humana experta mantuvo ventajas significativas. La creatividad estudiantil mostró resultados heterogéneos, indicando que la efectividad depende del contexto específico y el nivel de conocimiento previo del estudiante. En segundo lugar, los principales riesgos éticos documentados incluyen violaciones de privacidad por uso no autorizado de APIs, erosión de la integridad académica por dependencia excesiva, y perpetuación de inequidades cuando estudiantes con conocimiento limitado no obtienen beneficios. Las alucinaciones de LLMs representan un riesgo crítico para la precisión educativa, mientras que la ausencia de marcos éticos consistentes entre instituciones genera aplicación desigual de estándares. Los sesgos algorítmicos persisten en recomendaciones de aprendizaje, afectando particularmente a poblaciones vulnerables. En tercer lugar, las estrategias más exitosas combinaron metodologías pedagógicas estructuradas con capacidades específicas de LLMs. La integración con aprendizaje basado en problemas logró tasas de finalización del 86.4%, mientras que la ingeniería de prompts específicos mejoró la precisión 4.926 veces. Los sistemas híbridos instructor-IA superaron métodos automatizados puros, y el uso

moderado de funciones específicas alcanzó tasas de aprobación del 72.22%. El marco de ciclo de vida de desarrollo de prompts (PDLC) demostró efectividad en la formación estudiantil. Finalmente, los factores motivacionales identificados incluyen metas autotrascendentes como predictor principal de uso efectivo, personalización mediante alineación con estilos de aprendizaje, desarrollo de autonomía estudiantil, percepción de utilidad práctica, y provision de feedback inmediato. Los estudiantes orientados hacia beneficios colectivos mostraron mejores resultados colaborativos. La adaptación a marcos como VARK generó mejoras significativas en motivación, mientras que el apoyo emocional y la reducción de ansiedad emergieron como factores críticos para el bienestar estudiantil.

V. CONCLUSIONES

El objetivo principal de esta RSL demuestra que los LLMs constituyen ser herramientas educativas efectivas en dominios cognitivos específicos, superando significativamente a métodos convencionales en áreas que requieren retroalimentación inmediata, personalización y escalabilidad. La evidencia sistemática indica mejoras consistentes en rendimiento académico en disciplinas como criptografía cuántica, medicina y álgebra, con tasas de éxito del 86.4% en implementaciones estructuradas. No obstante, las limitaciones son evidentes en habilidades motoras y competencias que demandan supervisión humana experta directa. Además, los hallazgos revelan que la efectividad de los LLMs depende críticamente del contexto de implementación, nivel de conocimiento previo del estudiante, moderación en el uso, y adopción de marcos pedagógicos estructurados. El bienestar estudiantil se ve positivamente impactado cuando los LLMs proporcionan apoyo emocional y reducen la ansiedad académica, aunque persisten riesgos éticos significativos relacionados con privacidad, integridad académica y equidad educativa. Asimismo, las instituciones educativas deben desarrollar marcos éticos consistentes que aborden privacidad de datos, integridad académica y equidad antes de implementar LLMs. Se recomienda adoptar estrategias pedagógicas híbridas que integren LLMs con aprendizaje basado en problemas, priorizando el uso moderado y la ingeniería de prompts específicos. La formación docente debe incluir competencias en supervisión de sistemas LLM y desarrollo de habilidades de prompt engineering. Finalmente, las investigaciones futuras deben centrarse en el desarrollo de marcos regulatorios unificados para el uso ético de LLMs en educación, incluyendo protocolos de privacidad y prevención de sesgos. Se requieren estudios longitudinales que evalúen el impacto a largo plazo de los LLMs en el desarrollo de habilidades fundamentales y capacidades de resolución independiente. Es necesario investigar estrategias específicas para optimizar el uso de LLMs en diferentes tipos de competencias, especialmente en habilidades motoras y áreas donde la supervisión humana tradicional muestra superioridad. Los trabajos futuros deben explorar el desarrollo de sistemas adaptativos más sofisticados que personalicen la experiencia educativa según perfiles individuales de aprendizaje, y establecer métricas estandarizadas para evaluar la efectividad de diferentes implementaciones de LLMs en contextos educativos diversos. Finalmente, se requiere investigación sobre la integración efectiva de LLMs con tecnologías emergentes como realidad aumentada y robótica social para crear experiencias educativas inmersivas y completas.

REFERENCIAS

[1] R. Sajja, Y. Sermet, M. Cikmaz, D. Cwiertny, and I. Demir, "Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education," Information, vol. 15, no. 10, p. 596, 2024.

[2] A. Coban, D. Dzsotjan, S. Küchemann, J. Durst, J. Kuhn, and C. Hoyer, "AI support meets AR visualization for Alice and Bob: personalized learning based on individual ChatGPT feedback in an AR quantum cryptography experiment for physics lab courses," EPJ Quantum Technol., vol. 12, no. 1, p. 15, 2025.

[3] D. T. K. Ng, C. W. Tan, and J. K. L. Leung, "Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study," Br. J. Educ. Technol., vol. 55, no. 4, pp. 1328-1353, 2024.

[4] G. Polverini and B. Gregorcic, "How understanding large language models can inform the use of ChatGPT in physics education," Eur. J. Phys., vol. 45, no. 2, p. 025701, 2024.

[5] C. Li, W. Xing, Y. Song, and B. Lyu, "RICE AlgebraBot: Lessons learned from designing and developing responsible conversational AI using induction, concretization, and exemplification to support algebra learning," Comput. Educ. Artif. Intell., vol. 8, p. 100338, 2025.

[6] A. Borg, et al., "Virtual patient simulations using social robotics combined with large language models for clinical reasoning training in medical education: mixed methods study," J. Med. Internet Res., vol. 27, p. e63312, 2025.

[7] A. Yamamoto, M. Koda, H. Ogawa, T. Miyoshi, Y. Maeda, F. Otsuka, and H. Ino, "Enhancing Medical Interview Skills Through AI-Simulated Patient Interactions: Nonrandomized Controlled Trial," JMIR Med. Educ., vol. 10, no. 1, p. e58753, 2024.

[8] V. Kozov, B. Ivanova, K. Shoylekova, and M. Andreeva, "Analyzing the Impact of a Structured LLM Workshop in Different Education Levels," Appl. Sci., vol. 14, no. 14, 2024.

[9] R. S. Kusuma and T. Oktavia, "Leveraging AI-powered adaptive tutoring to address motivation challenges in university pedagogy," J. Logist. Inform. Serv. Sci., vol. 11, no. 8, pp. 42–60, 2024.

[10] R. Sajja, Y. Sermet, M. Cikmaz, D. Cwiertny, and I. Demir, "Artificial intelligence-enabled intelligent assistant for personalized and adaptive learning in higher education," Information, vol. 15, no. 10, p. 596, 2024.

[11] C. H. Lai and C. Y. Lin, "Analysis of Learning Behaviors and Outcomes for Students with Different Knowledge Levels: A Case Study of Intelligent Tutoring System for Coding and Learning (ITS-CAL)," Appl. Sci., vol. 15, no. 4, 2025.

[12] R. Azoulay, T. Hirst, and S. Reches, "Large Language Models in Computer Science Classrooms: Ethical Challenges and Strategic Solutions," Appl. Sci., vol. 15, no. 4, 2025.

[13] Y. Qiu, "The impact of llm hallucinations on motor skill learning: A case study in badminton," IEEE Access, 2024.

[14] S. Pozdniakov, et al., "Large language models meet user interfaces: The case of provisioning feedback," Comput. Educ. Artif. Intell., vol. 7, p. 100289, 2024.

[15] P. Y. W. Myint, S. L. Lo, and Y. Zhang, "Harnessing the power of AI-instructor collaborative grading approach: Topic-based effective grading for semi open-ended multipart questions," Comput. Educ. Artif. Intell., vol. 7, p. 100339, 2024.

[16] D. Scherbakov, N. Hubig, V. Jansari, A. Bakumenko, and L. A. Lenert, "The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review," arXiv preprint, 2024.

[17] J. L. Lieberum, et al., “Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review,” J. Clin. Epidemiol., vol. 181, p. 111746, 2025.
[18] C. H. Krag, et al., “Large language models for abstract screening in systematic-and scoping reviews: A diagnostic test accuracy study,” medRxiv, pp. 2024-10, 2024.
[19] E. Nasarian, R. Alizadehsani, U. R. Acharya, and K. L. Tsui, “Designing interpretable ML system to enhance trust in healthcare: A systematic review to proposed responsible clinician-AI-collaboration framework,” Inform. Fusion, p. 102412, 2024.
[20] X. Li, G. Han, B. Fang, and J. He, “Advancing the in-class dialogic quality: Developing an artificial intelligence-supported framework for classroom dialogue analysis,” The Asia-Pacific Educ. Res., vol. 34, no. 1, pp. 495-509, 2025.
[21] R. B. Toma and I. Yánez-Pérez, “Effects of ChatGPT use on undergraduate students’ creativity: a threat to creative thinking?,” Discover Artif. Intell., vol. 4, no. 1, p. 74, 2024.
[22] E. K. Nartey, “Guiding principles of generative AI for employability and learning in UK universities,” Cogent Educ., vol. 11, no. 1, p. 2357898, 2024.
[23] R. Maalek, “Integrating Generative Artificial Intelligence and Problem-Based Learning into the Digitization in Construction Curriculum,” Buildings, vol. 14, no. 11, 2024.
[24] L. Willey, B. J. White, and C. S. Deale, “Teaching AI in the college course: introducing the AI prompt development life cycle (PDLC),” Issues Inf. Syst., vol. 24, no. 2, 2023.
[25] C. Wang, et al., “Application of Large Language Models in Medical Training Evaluation—Using ChatGPT as a Standardized Patient: Multimetric Assessment,” J. Med. Internet Res., vol. 27, p. e59435, 2025.
[26] D. Huang and J. E. Katz, “GenAI Learning for Game Design: Both Prior Self-Transcendent Pursuit and Material Desire Contribute to a Positive Experience,” Big Data Cogn. Comput., vol. 9, no. 4, p. 78, 2025.
[27] P. Bitzenbauer, “ChatGPT in physics education: A pilot study on easy-to-implement activities,” Contemp. Educ. Technol., vol. 15, no. 3, p. ep430, 2023.