Empleo de la IA para la generación y evaluación de preguntas de examen en el curso de lingüística inglesa de pregrado*

Tina Xu Geng
https://orcid.org/0009-0004-2117-9617
Guangzhou College of Commerce,
China
xutianhua@gcc.edu.cn

Raquel Raya-Porcel
https://orcid.org/0009-0009-4158-3883
Durham New College,
Reino Unido
raquelraya@googlemail.com

Resumen

Este estudio explora la aplicación de la inteligencia artificial (IA) en la generación y evaluación de preguntas de examen para cursos de lingüística inglesa a nivel universitario. En particular, se compara el desempeño de dos modelos de IA: ChatGPT y DeepSeek, en la creación de preguntas justas, coherentes y variadas, así como su eficacia en la calificación de respuestas de tipo subjetivo. Mediante un enfoque metodológico mixto que incluye análisis cuantitativo de contenidos generados por IA y evaluación comparativa con criterios humanos, el estudio identifica ventajas clave en eficiencia, objetividad y adaptabilidad, junto con desafíos tales como la comprensión contextual limitada y dificultades para evaluar respuestas creativas. Los hallazgos indican que, si bien la IA puede apoyar significativamente la docencia, su aplicación exitosa requiere ajustes minuciosos, supervisión continua y consideraciones éticas. El estudio concluye con recomendaciones prácticas dirigidas a educadores para optimizar el uso de herramientas de evaluación basadas en IA, con el fin de mantener estándares académicos y mejorar el rendimiento estudiantil.

Palabras clave

Calificación de exámenes generados por IA; ChatGPT; DeepSeek; evaluación de la educación; examen; inteligencia artificial.

* Financiado por el Guangdong Provincial Education Science Planning Project: "Research on Eye Movement-Based English Reading Testing and Digital Evaluation Mechanism". Código del proyecto: 2024GXJK453

Recibido: 15/04/2025 | Enviado a pares: 03/11/2025 | Aceptado por pares: 25/11/2025 | Aprobado: 24/03/2026

DOI: 10.5294/edu.2026.29.1.6

Para citar este artículo / to reference this article / para citar este artigo

Xu, T. y Raya-Porcel, R. (2026). Empleo de la inteligencia artificial para la generación y evaluación de preguntas de examen en el curso de lingüística inglesa de pregrado. *Educación y Educadores, 29*(1), e2916. https://doi.org/10.5294/edu.2026.29.1.6

e-ISSN 2027-5358 | Educ.Educ. Vol. 29. No. 1 | e2916
Universidad de La Sabana | Facultad de Educación

Use of AI for Generating and Grading Exam Questions in an Undergraduate English Linguistics Course*

Abstract

This study explores the application of artificial intelligence (AI) in the generation and grading of exam questions for undergraduate English linguistics courses. Specifically, it compares the performance of two AI models, ChatGPT and DeepSeek, in creating fair, coherent, and varied questions, as well as their effectiveness in grading subjective responses. Using a mixed-methods approach that combines quantitative analysis of AI-generated content with comparative evaluation against human criteria, the study identifies key advantages in efficiency, objectivity, and adaptability, along with challenges such as limited contextual understanding and difficulties in assessing creative responses. The findings suggest that while AI can significantly support teaching, its successful implementation requires careful calibration, ongoing supervision, and ethical considerations. The study concludes with practical recommendations for educators to optimize the use of AI-based assessment tools in order to maintain academic standards and improve student performance.

Keywords

AI-generated exam grading; ChatGPT; DeepSeek; educational assessment; exams; artificial intelligence.

* Financed by the Guangdong Provincial Education Science Planning Project: "Research on Eye Movement-Based English Reading Testing and Digital Evaluation Mechanism". Grant number: 2024GXJK453

Uso da IA na geração e avaliação de questões de prova em cursos de graduação em linguística inglesa*

Resumo

Este estudo explora a aplicação da inteligência artificial (IA) na geração e avaliação de questões de prova em cursos universitários de linguística inglesa. Mais especificamente, compara o desempenho dos modelos ChatGPT e DeepSeek na elaboração de questões equilibradas, coerentes e diversificadas, bem como na avaliação de respostas discursivas. Utilizando uma abordagem metodológica mista, que inclui análise quantitativa do conteúdo gerado por IA e comparação com avaliação humana, o estudo identifica principais vantagens relacionadas à eficiência, objetividade e adaptabilidade, além de desafios como limitações na compreensão contextual e dificuldades na avaliação de respostas criativas. Os resultados indicam que, embora a IA possa apoiar significativamente os processos avaliativos, sua implementação bem-sucedida requer ajustes criteriosos, supervisão contínua e considerações éticas. O estudo conclui com recomendações práticas para que educadores otimizem o uso de ferramentas avaliativas baseadas em IA, a fim de manter os padrões acadêmicos e melhorar o desempenho estudantil.

Palavras-chave

Avaliação de provas geradas por IA; ChatGPT; DeepSeek; avaliação educacional; elaboração de provas; inteligência artificial.

* Financiado pelo Guangdong Provincial Education Science Planning Project: "Research on Eye Movement-Based English Reading Testing and Digital Evaluation Mechanism". Código do projeto: 2024GXJK453

A través de un análisis integral de las tecnologías actuales de inteligencia artificial (IA) y de una investigación empírica sobre sus aplicaciones pedagógicas, este artículo explora la integración de la IA en un curso de Introducción a la Lingüística Inglesa. Específicamente, examina el uso de dos modelos de IA generativa (IAGen): ChatGPT y DeepSeek, en la creación de contenido didáctico para el capítulo 5 sobre semántica; el diseño de preguntas de evaluación y la calificación de respuestas de examen. La investigación se centra en tres aspectos clave: primero, si el contenido didáctico generado por IA cumple con los requisitos establecidos por el plan de estudios y si los modelos de IA pueden producir preguntas de examen que reflejen con precisión el contenido del curso y la efectividad de la enseñanza; segundo, si IA puede proporcionar una puntuación justa y coherente; y tercero, las diferencias en el contenido generado y la precisión entre los dos modelos de IA generativa.

Más allá de evaluar la efectividad de la IA en la enseñanza, este trabajo busca establecer un marco de evaluación holístico, ofreciendo datos y análisis multimodales para fines pedagógicos. Los hallazgos proporcionarán una referencia para educadores e instituciones que deseen implementar estrategias de evaluación impulsadas por IA, abordando los desafíos en la creación de exámenes y la entrega de retroalimentación, mientras se mejora la experiencia educativa en general.

Estado del arte

En diciembre de 2022, el lanzamiento del chatbot IAGen ChatGPT-3 de OpenAI atrajo la atención mundial en el ámbito educativo. Investigadores identificaron diversas aplicaciones potenciales de la IA en el campo educativo, lo que dio lugar a una disciplina emergente: la inteligencia artificial en la educación (AIEd). Con el rápido avance de la tecnología de IA en los últimos dos años, sus aplicaciones en la educación han ganado impulso y demostrado un potencial significativo en la generación de contenido didáctico, el diseño de preguntas de examen y la corrección de evaluaciones. Los académicos han explorado ampliamente si la IAGen puede reemplazar a los docentes humanos (Leiker *et al.*, 2023; Cooper, 2023). En enero de 2025, el surgimiento de DeepSeek-R1 marcó un avance revolucionario y generó un notable impacto a nivel global debido a su código abierto, menores costos operativos y rendimiento superior al de ChatGPT. Este avance impulsó la investigación y práctica educativa basada en IA a buscar nuevas metas. A continuación, se analizan cuatro aspectos clave que ilustran las tendencias actuales de investigación y los avances en el campo de la IA en la educación (AIEd): 1) aplicaciones de la IA en la generación de contenido didáctico, 2) en la generación y evaluación de preguntas de examen, 3) los retos y oportunidades de la IA en las evaluaciones y 4) la comparación entre ChatGPT y DeepSeek.

Aplicaciones de la IA en la generación de contenido didáctico

El uso de IAGen para crear contenido didáctico se ha convertido en un área de investigación cada vez más destacada. Investigadores han analizado las diferencias entre contenido generado por IA y por humanos a nivel global (Weber-Wulff *et al.*, 2023; Rana *et al.*, 2024). Markowitz (2024, p. 63) defienden que, si bien los textos generados por IA son cada vez más indistinguibles de los escritos por humanos, persisten diferencias estilísticas notables: los textos de IA son más analíticos y descriptivos, más cargados emocionalmente, pero menos legibles que los textos humanos. Además, Chu y Liu (2024) encontraron que, en comparación con el contenido humano, el material generado por IA reduce eficazmente las respuestas críticas y promueve respuestas emocionales consistentes con la narrativa propuesta. Esta característica puede aumentar su atractivo y mejorar la experiencia de aprendizaje. Sin embargo, Chu y Liu (2023) también observaron que, cuando los estudiantes saben que el contenido es generado por IA, se produce el efecto contrario: disminuye la par-

ticipación, aumentan las objeciones y la efectividad de la enseñanza se reduce.

En cuanto a los resultados de aprendizaje, algunos estudios sugieren que el contenido generado por IA puede ser más eficaz que el generado por humanos en ciertos contextos (Leddo, 2021; Zhang *et al.*, 2024). Su estilo analítico y descriptivo puede mejorar la retención de hechos (Markowitz *et al.*, 2024) y sus narrativas, que generan menos contraargumentos (Chu y Liu, 2023), podrían facilitar una mejor internalización del conocimiento. No obstante, estos beneficios pueden verse contrarrestados por la desconfianza humana con el contenido de origen IA, especialmente cuando se informa de su origen IAGen. Los estudiantes están más acostumbrados al contenido humano, que incluye redundancias menores, saltos lógicos y errores, mientras que el contenido de IA puede parecer demasiado pulido, monótono y artificialmente perfecto. A pesar del potencial de IAGen en la generación de contenido didáctico, siguen siendo necesarias mejoras en la legibilidad y resonancia emocional.

Aplicaciones de la IA en la generación y evaluación de preguntas de examen

En lo que respecta a la resolución de problemas, los estudios han demostrado que ChatGPT puede obtener puntuaciones altas (entre 80 % y 90 %) en diversos exámenes, incluidos, pero no limitados, en física (Yeadon *et al.*, 2024), medicina (Kung *et al.*, 2023), derecho (Choi, 2024), cirugía (Gencer y Aydin, 2023) e investigación de operaciones (Farazouli *et al.*, 2024). Dado que los modelos de IA dan respuestas únicas en cada ocasión, los software de detección de plagio existentes tienen dificultades para detectar el uso de IA; además, los docentes también encuentran complicado distinguir las respuestas generadas por ChatGPT de las escritas por estudiantes (Farazouli *et al.*, 2024). Asimismo, algunos estudios han investigado la precisión de la IA en la resolución de exámenes. Por ejemplo, Xu y Zhou (2024) evaluaron cuantitativamente el rendimiento de ChatGPT-4 en preguntas anteriores del examen nacional de calificación para supervisores de enfermería, con una tasa de precisión general del 81 %. De manera similar, Sun (2024) encontró que ChatGPT obtuvo un 82 % de aciertos en preguntas de geografía del examen de ingreso universitario de la provincia de Jiangsu de 2023.

En cuanto al diseño de preguntas de examen, tradicionalmente la calificación de las mismas ha sido realizada manualmente por los docentes, un proceso que, además de ser laborioso y demandante, también es susceptible de contener juicios subjetivos. Los modelos de IA abren nuevas posibilidades para correcciones de exámenes automatizadas. Mabrito *et al.* (2025) aplicaron ChatGPT para evaluar las respuestas de seis estudiantes a 10 preguntas basadas en casos y observaron que los resultados fueron comparables a los de un docente humano. Aunque se observaron discrepancias en el 65 %-80 % de los casos, estas diferencias se consideraron dentro de un rango aceptable, lo que indica una tendencia de ChatGPT a asignar puntuaciones más altas que las que dan los profesores humanos. Alers *et al.* (2024) evaluaron 105 exámenes en los que se les solicitó que analizaran críticamente una investigación y hallaron una diferencia del 70,5 % entre las calificaciones de ChatGPT y las de los docentes, aunque las puntuaciones parecían correlacionarse. Pinto *et al.* (2023) encargaron a ChatGPT corregir y calificar respuestas de un programa de formación de una empresa de software, con las puntuaciones finales aprobadas por el formador de la empresa. Otros estudios han explorado la capacidad de ChatGPT para procesar imágenes con el fin de evaluar materiales producidos por estudiantes (Lee y Zhai, 2024). Además, ChatGPT ha demostrado ser capaz de proporcionar retroalimentación (*feedback*) clara, coherente y personalizada en planificación de proyectos de estudiantes de ciencias (Dai *et al.*, 2023). Para preguntas objetivas (Bewersdorff *et al.*, 2023), la IA puede identificar con precisión las respuestas de los estudiantes, compararlas con las respuestas correctas y asignar puntuaciones. Para preguntas subjetivas, la IA emplea análisis semántico,

análisis sintáctico y otras técnicas para realizar evaluaciones preliminares, lo cual mejora la eficiencia y la equidad en la calificación.

Retos y oportunidades de la IA en las evaluaciones

Antes del surgimiento de IAGen, los sistemas de evaluación automatizada requerían un entrenamiento extenso basado en respuestas previamente corregidas de forma manual, para desarrollar modelos de puntuación (Zupanc y Bosnić, 2020). Una característica distintiva de los modelos de lenguaje de gran tamaño (LLMs), como ChatGPT, es su capacidad para evaluar exámenes sin entrenamiento previo ni instrucciones explícitas sobre los criterios de evaluación. Sin embargo, persisten limitaciones. Deng *et al.* (2025) utilizaron IA para identificar errores gramaticales en textos escritos y descubrieron que ChatGPT tenía dificultades para reconocer dichos errores. Además, Daun y Brings (2023) discutieron el uso de IAGen en la enseñanza a estudiantes de ingeniería de software y destacaron los posibles riesgos de que los alumnos dependan de esta tecnología en sus estudios.

Comparación entre ChatGPT y DeepSeek

En el primer mes tras el lanzamiento de DeepSeek, se publicaron dos artículos en Scopus comparando ChatGPT con DeepSeek, centrados en aplicaciones empíricas de la IA en el ámbito de la salud (Peng *et al.*, 2025; Kayaalp *et al.*, 2025). Al 6 de marzo de 2025, una búsqueda de "DeepSeek" en la base de datos China National Knowledge Infrastructure (CNKI) arrojó 28 artículos, de los cuales 12 discutían las ventajas y aplicaciones futuras de DeepSeek, seis analizaban sus impactos sociales y dilemas éticos y diez exploraban sus perspectivas futuras en campos específicos, entre ellos, seis abordaban posibles aplicaciones pedagógicas, aunque ninguno incluía investigaciones empíricas ni evaluaciones comparativas.

En comparación con ChatGPT, la naturaleza de código abierto y bajo costo de DeepSeek ofrece mayor flexibilidad para aplicaciones pedagógicas. Aunque ChatGPT sobresale en la creación de lenguaje y tareas conversacionales, sus altos costos operativos limitan su uso generalizado. En contraste, el modelo de código abierto de DeepSeek reduce las barreras técnicas de entrada y promueve un acceso más equitativo a los recursos educativos.

Métodos de investigación

Este estudio seleccionó 81 exámenes completos de las clases 2203 y 2204 de la carrera de Inglés en la Facultad de Lenguas Extranjeras del Guangzhou College of Commerce. El examen se realizó en diciembre de 2024, en inglés, y evaluaba el capítulo 5 sobre Semántica del curso Introducción a la Lingüística Inglesa. Tras el examen, siguiendo las directrices de la universidad, las copias físicas fueron corregidas manualmente por su profesor, mientras que las copias digitales escaneadas fueron evaluadas de manera independiente por dos modelos de IA: ChatGPT y DeepSeek. El curso es obligatorio, con un valor de dos créditos, y está alineado con las directrices nacionales chinas de educación superior. La evaluación utiliza la escala convencional de 100 puntos. Para la corrección por IA, se usaron indicaciones específicas adaptadas a cada rol con el fin de asegurar la coherencia.

Para evitar que los modelos de IA se vieran influenciados por correcciones anteriores, se abrió un nuevo chat para cada examen. Se pidió tanto a ChatGPT como a DeepSeek que calificaran cada examen tres veces, utilizando el promedio aritmético (redondeado al número entero más cercano) de los tres intentos como calificación final. Este estudio generó un total de 498 calificaciones con IA. El análisis comparativo tiene dos enfoques: comparaciones horizontales entre ChatGPT y DeepSeek y comparaciones verticales entre las calificaciones de la IA y las del docente humano. Se utilizaron como métricas de equivalencia el Kappa de Cohen y el Kappa ponderado

linealmente de Cohen, para evaluar las diferencias entre las calificaciones generadas por la IA y las del profesor (Falchikov y Goldfinch, 2000; Sadler y Good, 2006). El Kappa de Cohen mide el grado de acuerdo entre dos promedios, así como la probabilidad de coincidencia por azar. El Kappa ponderado de Cohen extiende esta métrica al considerar la magnitud de los desacuerdos. Los niveles de concordancia se clasifican de la siguiente forma: entre 0 y 1, < 0,2 indica un acuerdo leve; 0,21-0,4, acuerdo razonable; 0,41-0,6, moderado; 0,61-0,8, significativo; >0,81, acuerdo casi perfecto (Viera y Garrett, 2008).

Además, se implementó un cuestionario entre los 81 estudiantes –con dos que no participaron, para 79 respuestas válidas– con el fin de explorar la aceptación del uso de IA en enseñanza y evaluación. También se realizaron entrevistas con el profesor para conocer su actitud, expectativas y retroalimentación respecto al uso de IA en la elaboración y corrección de exámenes.

El presente estudio adopta un diseño mixto convergente de tipo paralelo, en el cual los datos cuantitativos y cualitativos se recopilan y analizan de manera simultánea dentro de la misma fase, lo que permite contextualizar las tendencias estadísticas en la calificación a partir de las percepciones del profesorado y del alumnado. El componente cuantitativo se compone de: a) la comparación estadística de las calificaciones de exámenes asignadas por docentes humanos, ChatGPT y DeepSeek mediante los coeficientes Kappa de Cohen y Kappa ponderado lineal, y b) el análisis descriptivo de los cuestionarios respondidos por 79 estudiantes. Por su parte, el componente cualitativo incluye entrevistas semiestructuradas realizadas al docente responsable del curso, con el fin de explorar sus percepciones, expectativas y valoraciones sobre el uso de la inteligencia artificial en el diseño y la corrección de exámenes. La integración de ambos enfoques se llevó a cabo en la fase de interpretación, lo que permitió obtener una comprensión integral de la eficacia y la aceptación de la IA en la evaluación educativa.

El uso de los datos del alumnado se realizó siguiendo estrictos principios éticos. Todos los participantes fueron informados sobre los objetivos del estudio y otorgaron su consentimiento de manera voluntaria. Con el fin de proteger la privacidad, todas las copias de los exámenes fueron anonimizadas antes de ser procesadas por los modelos de IA, para garantizar que no se compartiera información personal identificable con servidores externos. Asimismo, el uso de herramientas de inteligencia artificial se limitó exclusivamente a fines de investigación y mejora pedagógica, sin influir en las calificaciones oficiales del alumnado. El estudio respeta los principios de confidencialidad, transparencia y uso responsable de la tecnología en contextos educativos.

El cuestionario utilizado para analizar la aceptación del alumnado (Tabla 6) fue sometido a una fase piloto con un pequeño grupo de estudiantes de una cohorte previa, con el objetivo de garantizar la claridad y pertinencia de los ítems. Además, su diseño se fundamentó en estudios previos sobre la percepción del uso de la inteligencia artificial en la educación, lo que respalda su validez de contenido.

Con el fin de reforzar la coherencia metodológica, los resultados cuantitativos derivados del análisis del Kappa de Cohen (ver abajo en "Calificación") fueron triangulados de manera sistemática con la información cualitativa obtenida a partir de las entrevistas al profesorado. Esta integración permite comprender con mayor profundidad las razones por las cuales los modelos de IA pueden diferir del juicio humano en contextos creativos o subjetivos y contribuye a cerrar la brecha entre la correlación estadística y la realidad pedagógica.

Análisis y discusión

Primero se compara el desempeño de ChatGPT y DeepSeek en la creación del contenido de la presentación (PPT), enfocándose en la cobertura curricular y la profundidad del contenido. Enseguida se discute la generación de preguntas de examen por

parte de la IA, evaluando su alineación con los criterios del programa y comparando la precisión de las preguntas generadas por ambos modelos. Después se examinan los resultados de la corrección con IA, evaluando la consistencia de los resultados otorgados por ChatGPT, DeepSeek y el docente humano, así como la equidad y fiabilidad de la calificación automática. Finalmente, se analizan las respuestas al cuestionario, explorando la aceptación y percepciones del alumnado sobre la enseñanza y evaluación generadas por IA.

Generación de presentaciones didácticas (PPT)

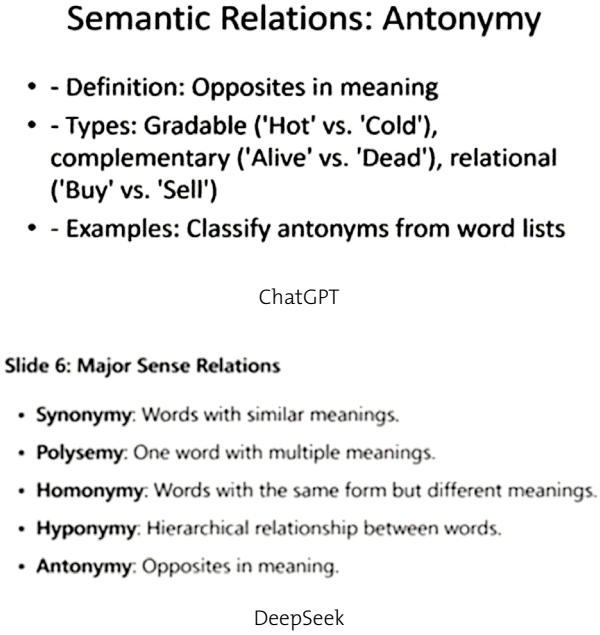
Se utilizó IAGen para generar presentaciones (PPT) proporcionando a los modelos de IA las siguientes indicaciones (*prompts*) relevantes:

> *Soy profesor universitario de Lingüística Inglesa. La próxima semana, voy a enseñar el capítulo 5: Semántica. El libro de texto utilizado es* 新编简明英语语言学教程, *ISBN: 9787810804028. ¿Puedes generar el contenido de la lección basándote en el libro que he adjuntado? Este capítulo tomará aproximadamente 3 horas. ¿Podrías también generar el contenido de la presentación en PowerPoint para esta lección?*

ChatGPT generó una presentación de 14 diapositivas que cubría de manera integral los contenidos clave del capítulo. Sin embargo, en cuanto a las explicaciones conceptuales, las discusiones sobre el conductismo y la sinonimia estaban poco desarrolladas, especialmente la descripción de los elementos esenciales de estos temas. En cuanto a las relaciones entre oraciones, esta sección estaba relativamente poco desarrollada y no logró explicar de manera exhaustiva las relaciones entre frases y sus ejemplos.

Por otro lado, DeepSeek produjo una presentación de 12 diapositivas que abarcaba de manera integral todos los conceptos. No obstante, su primera respuesta fue demasiado general y superficial; solo tras múltiples solicitudes de ejemplos y mayor detalle logró generar diapositivas más adecuadas. La Figura 1 muestra las diapositivas producidas por ChatGPT y DeepSeek sobre el tema de las relaciones de sentido, utilizando exactamente la misma indicación o *prompt*. Como se ve en la Figura 1, DeepSeek no logró generar en su primer intento diapositivas aptas para la enseñanza que incluyeran ejemplos específicos o detalles suficientes y ofreció únicamente un resumen altamente generalizado.

Figura 1. Presentaciones didácticas generadas por ChatGPT y DeepSeek



Fuente: elaboración propia.

En comparación con las diapositivas generadas por IA, las diapositivas creadas por el docente presentan las siguientes diferencias notables, también reflejadas en los resultados del cuestionario:

Paleta de colores más rica. Las diapositivas generadas por humanos tienen una mayor variedad cromática, lo que crea un efecto visual más atractivo que mejora la atención y participación del alumnado. Esta diversidad de colores puede, en cierta medida, reducir la fatiga de aprendizaje y mejorar el ambiente del aula.

Mayor participación estudiantil. Las diapositivas diseñadas por el docente incorporan ejemplos vívidos y del mundo real que hacen que los conceptos semánticos abstractos sean más comprensibles y relevantes. Este enfoque estimula el interés del alumnado y lo alienta a participar activamente en las discusiones.

Ejemplos vinculados a la vida estudiantil. Los ejemplos generados por el docente están más relacionados con la vida cotidiana del alumnado y escenarios del mundo real, lo que mejora la relevancia del contenido. Esta contextualización potencia la comprensión y retención de los contenidos lingüísticos y ayuda a los estudiantes a aplicar mejor la teoría semántica en su uso real del idioma inglés.

Generación de preguntas de examen

La IA puede generar preguntas de examen que se alinean con los requisitos del programa de estudios, al cubrir eficazmente conceptos clave y adaptarse a distintos niveles de dificultad y formatos. Además, la IA también puede proporcionar respuestas. La Tabla 1 compara las tasas de precisión de dos modelos de IA. Ambos exámenes consisten exclusivamente en preguntas de opción múltiple. El examen 1 fue generado por ChatGPT, mientras que el examen 2 es el oficial del capítulo 5 utilizado en el curso; ambos contienen 15 preguntas de dificultad similar. Como muestra la Tabla 1, DeepSeek alcanzó la mayor precisión, con un 93 %, seguido por ChatGPT, con un 90 %.

Tabla 1. Comparación de errores de los modelos de IA en los resultados de las pruebas

	DeepSeek	ChatGPT
Examen 1 (generado por ChatGPT)	2 errores	2 errores
Examen 2 (generado por humanos)	0 errores	1 error
Tasa de precisión promedio	93 %	90 %

Fuente: elaboración propia.

Como se muestra en la Tabla 2, DeepSeek respondió incorrectamente a las preguntas 7 y 14 del examen 1 y ChatGPT falló en las preguntas 7 y 15 del mismo examen. Las causas principales de estos errores radican en que la IA aún no distingue con precisión conceptos lingüísticos complejos y detallados. Además, la IA no es suficientemente sensible a las características estilísticas y diferencias sutiles en la elección léxica, ni posee una comprensión profunda de la versatilidad y dependencia contextual de los fenómenos del lenguaje. Estos errores reflejan las limitaciones actuales de los sistemas de IA para comprender las características semánticas y pragmáticas más sofisticadas del lenguaje natural, especialmente en preguntas que requieren una aplicación crítica de conocimientos lingüísticos académicos.

Calificación

Los datos para esta parte del estudio provienen de los resultados del examen del capítulo 5 del curso Introducción a la Lingüística Inglesa, corregidos de manera independiente por profesores humanos, ChatGPT y DeepSeek. Se analizaron las calificaciones de un total de 81 estudiantes, utilizando una escala convencional de 100 puntos. Para evaluar la fiabilidad entre evaluadores, se emplearon los coeficientes Kappa de Cohen y Kappa ponderado lineal. Estas métricas miden eficazmente el grado de acuerdo entre evaluadores (como en calificaciones) y consideran las discrepancias entre ellos. En el proceso de evaluación automatizada, se utilizó la siguiente indicación:

> *Soy profesor de Lingüística Inglesa. Por favor, califica este examen (fotos adjuntas) sobre una escala del 100%. El examen consta de cuatro preguntas:*
>
> 1. *Analiza la relación entre "sentido" y "referencia" con ejemplos. Proporciona un ejemplo en el que dos expresiones tengan la misma referencia, pero sentidos distintos. Explica la importancia de esta distinción.*

Tabla 2. Preguntas respondidas incorrectamente por IA

	Pregunta 3 (Examen 2)	**Pregunta 7 (Examen 1)**	**Pregunta 14 (Examen 1)**
Enunciado	X: They are going to have another baby. Y: They have a child. ¿Cuál es la relación entre X e Y?	"Baggage" y "Luggage" son ejemplos de:	Los homófonos suelen utilizarse para crear juegos de palabras con:
Respuesta correcta	D	C	D
Respuesta incorrecta (IA)	C	D	A

Fuente: elaboración propia.

2. *Explica cómo el contexto afecta el significado de una palabra con ejemplos específicos. Discute el papel del contexto situacional y lingüístico en la interpretación del significado. Ejemplo: a) "The seal was swimming near the shore." b) "The seal on the document is authentic."*

3. *Clasifica los siguientes antónimos y justifica tu clasificación. Identifica si son antónimos graduales, complementarios o relacionales, y explica por qué: a) "Big" vs. "Small"; b) "Alive" vs. "Dead"; c) "Doctor" vs. "Patient."*

4. *Utilizando el análisis componencial, descompón los significados de "boy," "man," y "girl" en rasgos semánticos. Muestra los rasgos compartidos y las diferencias, y comenta cómo este análisis ayuda a entender su relación.*

El uso de IAGen demuestra un valor único, pero también limitaciones. Por un lado, la IA puede procesar rápidamente grandes volúmenes de respuestas y proporcionar calificaciones y retroalimentación preliminar, lo cual es fundamental en evaluaciones a gran escala. Sin embargo, existen discrepancias en las calificaciones de preguntas subjetivas entre la IA y los docentes humanos.

Como se muestra en la Tabla 3, en cuanto al rango de puntuación, DeepSeek presentó el rango más amplio (20-100), lo que indica gran variabilidad; ChatGPT tuvo el rango más estrecho (63-98), reflejo de puntuaciones más concentradas; el rango de los docentes fue moderado (46-86), sin valores extremos, y las calificaciones fueron estables. En promedio, ChatGPT otorgó la calificación más alta (87,25), por encima de DeepSeek (72,62) y de los docentes (73,30), lo que sugiere cierta indulgencia en la puntuación. La desviación estándar de DeepSeek (17,667) fue la más alta, lo que indica mayor dispersión; ChatGPT presentó la más baja (5,276), señalando escasa variación; y los docentes humanos una media razonable (7,958), con puntuaciones coherentes y estables.

Tabla 3. Comparación de calificaciones en preguntas subjetivas (IA vs. docentes)

	Mín.	**Máx.**	**Suma**	**Promedio**	**Desviación estándar**
DeepSeek	20	100	5.882	72,62	17,667
ChatGPT	63	98	7.067	87,25	5,276
Docentes	46	86	5.937	73,30	7,958

Fuente: elaboración propia.

Realizando un análisis discreto de los datos, dado que los coeficientes Kappa de Cohen y Kappa lineal ponderado de Cohen son adecuados para datos categóricos, este estudio primero dividió los datos continuos de puntuación en cinco niveles (A, B, C, D, F); las reglas específicas son las siguientes: A: 90-100 puntos; B: 80-89 puntos; C: 70-79 puntos; D: 60-69 puntos; F: 0-59 puntos. Los datos categorizados se utilizan para los cálculos posteriores del coeficiente Kappa. El coeficiente Kappa de Cohen se utiliza para medir el grado de acuerdo entre dos evaluadores y su valor varía entre -1 y 1. La interpretación de los valores de Kappa es la siguiente: < 0: peor que la consistencia aleatoria; 0-0,2: consistencia leve; 0,21-0,4: consistencia general; 0,41-0,6: consistencia moderada; 0,61-0,8: consistencia significativa; 0,81-1: consistencia casi perfecta. A continuación, la Tabla 4 muestra la puntuación comparativa después de la organización y combinación.

El Kappa lineal ponderado de Cohen, que se muestra en la Tabla 5, tiene en cuenta la magnitud de las discrepancias entre las calificaciones, por lo que es adecuado para datos categóricos ordenados (como las notas). Su rango de valores es el mismo que el del Kappa de Cohen, pero ofrece una representación más precisa del grado de concordancia entre evaluadores.

Tabla 4. Comparaciones de los valores de Kappa

	Resultados Kappa
Humano vs. ChatGPT	Al calcular la consistencia de las puntuaciones entre los docentes humanos y ChatGPT, el valor de Kappa de Cohen es 0,65, lo que indica que existe una consistencia significativa entre ambos. Esto muestra que ChatGPT puede simular los criterios de calificación de los docentes humanos en esta tarea, aunque aún existen diferencias menores.
Humano vs. DeepSeek	El valor de Kappa de Cohen para la consistencia de las puntuaciones entre los docentes humanos y DeepSeek es 0,72, lo que indica que la consistencia entre ambos es mayor que la de ChatGPT.
ChatGPT vs. DeepSeek	El valor de Kappa de Cohen entre ChatGPT y DeepSeek es 0,58, lo que indica una consistencia moderada entre ambos. Esto muestra que los dos modelos de IA presentan algunas diferencias en sus criterios de calificación, aunque comparten ciertas similitudes en términos generales.

Fuente: elaboración propia.

Tabla 5. Comparación de los valores de Kappa ponderado

	Resultados de Kappa Ponderado
Docentes humanos vs. ChatGPT	El valor de Kappa lineal ponderado entre los docentes humanos y ChatGPT es 0,70, lo que indica un acuerdo significativo entre ambos. En coherencia con el resultado del Kappa de Cohen, las calificaciones de ChatGPT son similares a las de los docentes humanos.
Docentes humanos vs. DeepSeek	El valor de Kappa Lineal Ponderado entre los docentes humanos y DeepSeek es 0,75, lo que confirma aún más que las calificaciones de DeepSeek se asemejan más a las de los docentes humanos.
ChatGPT vs. DeepSeek	El valor de Kappa lineal ponderado entre ChatGPT y DeepSeek es 0,62, lo que indica un acuerdo significativo entre ambos modelos.

Fuente: elaboración propia.

La consistencia entre las calificaciones otorgadas por docentes humanos y la inteligencia artificial se refleja en el hecho de que la concordancia entre los educadores humanos y las puntuaciones de ChatGPT y DeepSeek alcanzó un nivel significativo (Kappa > 0,60), lo que indica que la IA generativa tiene un alto potencial de aplicación en la tarea de calificación del curso Introducción a la Lingüística Inglesa. Las puntuaciones de DeepSeek son más coherentes con las de los docentes humanos (Kappa = 0,72), lo que podría deberse a que sus criterios de calificación se asemejan más a los hábitos de evaluación humanos. Las diferencias entre los modelos de calificación de IA se evidencian principalmente en el hecho de que la consistencia entre las puntuaciones de ChatGPT y DeepSeek es de nivel moderado (Kappa = 0,58), indicativo de que existen ciertas discrepancias en los criterios de evaluación de ambos modelos. Estas diferencias pueden estar relacionadas con variaciones en sus datos de entrenamiento y en sus marcos algorítmicos. Investigaciones futuras podrían explorar cómo optimizar los modelos de calificación de IA para mejorar su coherencia y precisión.

Mediante el análisis de los valores de Kappa de Cohen y Kappa lineal ponderado de Cohen, este estudio descubrió una fuerte correlación entre las calificaciones de los docentes humanos y las de ChatGPT y DeepSeek en la tarea de calificación de trabajos. El nivel de acuerdo entre las calificaciones proporcionadas por la IA y los educadores humanos fue superior al 85 %, lo que indica que la calificación con IA puede ser confiable en muchos casos. Sin embargo, la IA puede enfrentar dificultades al interpretar las respuestas de los estudiantes, especialmente en preguntas abiertas, ya que le resulta más complicado captar la profundidad del entendimiento o los aportes únicos expresados en los escritos de los alumnos. Los datos muestran que, en lo que respecta a soluciones creativas, la precisión de las evaluaciones de la IA desciende a aproximadamente 70%, especialmente en situaciones que requieren pensamiento crítico y analítico. Por lo tanto, al calificar trabajos, los docentes pueden utilizar la IA para asignar calificaciones preliminares y luego realizar una revisión más detallada basada en la retroalimentación proporcionada por la IA. De esta manera, la IA generativa no solo puede mejorar la eficiencia en la enseñanza, sino también contribuir a una evaluación más justa y completa. Las investigaciones futuras pueden centrarse en cómo optimizar los algoritmos de IA para mejorar su capacidad de comprensión de respuestas complejas y así servir mejor al ámbito educativo.

Cuestionarios y entrevistas

Con el fin de explorar la aceptación del uso de inteligencia artificial en la educación por parte del alumnado, este estudio llevó a cabo cuestionarios y entrevistas. Los métodos específicos se detallaron arriba. A continuación, se presentan los principales hallazgos, y los resultados del cuestionario se muestran en la Tabla 6.

Tabla 6. Resultados del cuestionario

Preguntas del cuestionario	Respuestas mayoritarias	Respuestas minoritarias
1. ¿Puedes predecir qué prueba fue generada por IA?	Sí (58 votos, 73 %).	No (21 votos, 27 %).
2. ¿Tus predicciones fueron acertadas?	Incorrectas (41 votos, 52 %).	Correctas (38 votos, 48 %).
3. ¿Cómo compara la dificultad de las preguntas generadas por ChatGPT con las diseñadas por el docente?	Comparable (60 votos, 76 %).	Más difíciles (8 votos, 10 %); más fáciles (11 votos, 14 %).

Preguntas del cuestionario	Respuestas mayoritarias	Respuestas minoritarias
4. Preguntas generadas por ChatGPT vs. preguntas diseñadas por el docente.	Sin diferencias notables (45 votos, 58 %).	Más claras (15 votos, 19 %); mejor alineadas con el libro de texto (6 votos, 8 %); más difíciles de entender (11 votos, 14 %).
5. ¿Las preguntas generadas por ChatGPT mejoran tu comprensión de la lingüística inglesa?	Neutral (53 votos, 67 %).	Útiles (23 votos, 29 %); poco útiles (3 votos, 4 %).
6. Tu experiencia al completar preguntas generadas por ChatGPT.	Razonablemente complicadas (65 votos, 82 %).	Fáciles e interesantes (8 votos, 10 %); demasiado complejas, redujeron el interés (6 votos, 8 %).
7. ¿Estás dispuesto/a a usar ChatGPT para otras asignaturas?	Dispuesto/a (55 votos, 70 %).	Indeciso/a (21 votos, 27 %); no dispuesto/a (3 votos, 4 %).

Fuente: elaboración propia.

1. Reconocimiento inexacto de las preguntas generadas por IA. Los resultados del cuestionario muestran que el 73 % (58 estudiantes) afirmaron poder identificar las preguntas generadas por IA, mientras que el 27 % (21 estudiantes) dijeron no poder hacerlo. Sin embargo, al verificar sus respuestas, solo el 48 % (38 estudiantes) predijeron correctamente el origen de las preguntas, mientras que el 52 % (41 estudiantes) se equivocaron. Esto indica que, aunque la mayoría cree poder diferenciar las preguntas de IA de las humanas, sus juicios no son siempre acertados. Además, refleja que el estilo lingüístico y la lógica de las preguntas generadas por IA ya son comparables a las formuladas por docentes humanos.

2. Calidad aceptable de las preguntas de examen generadas por IA. En cuanto a la dificultad de las preguntas generadas por ChatGPT, el 76 % (60 estudiantes) consideraron que era equivalente a la de las preguntas diseñadas por el profesor, un 10 % (8 estudiantes) las consideraron más difíciles y a un 14 % (11 estudiantes) les parecieron más fáciles. Además, en cuanto a la calidad general de estas preguntas, el 58 % (45 estudiantes) opinó que "no había diferencias significativas" respecto a las preguntas docentes, el 19 % (15 estudiantes) las encontró "más claras", el 8 % (6 estudiantes) dijo que estaban "mejor alineadas con el libro de texto" y el 14 % (11 estudiantes) las encontró "más difíciles de entender".

Estos resultados demuestran que, para la mayoría del alumnado, las preguntas generadas por IA son aceptables en términos de calidad e incluso, en algunos casos, superiores. Sin embargo, una parte del alumnado sigue considerando que estas preguntas pueden ser poco claras o confusas, lo cual sugiere que las IA deben mejorar su capacidad de alinearse con los objetivos curriculares y reducir ambigüedades innecesarias.

3. Experiencia positiva del aprendizaje asistido por IA. Cuando se les preguntó si las preguntas de examen generadas por IA ayudaban a mejorar su comprensión de la lingüística inglesa, el 67 % (53 estudiantes) respondió que el impacto fue "neutral", el 29 % (23 estudiantes) dijo que eran útiles y solo el 4 % (3 estudiantes) las consideró poco útiles. Esto indica que la mayoría del alumnado mantiene una actitud neutral o positiva con respecto al rol complementario de la IA en el aprendizaje, aunque aún se requiere investigar más a fondo sus efectos a largo plazo.

Sobre el uso de la IA en otras asignaturas, el 70 % (55 estudiantes) se mostraron dispuestos a utilizar ChatGPT como herramienta de apoyo en el estudio,

el 27 % (21 estudiantes) estaban indecisos y solo el 4 % (3 estudiantes) se mostraron reacios. Esto sugiere que existe una amplia aceptación del aprendizaje asistido por IA, lo que refleja su gran potencial en el futuro de la educación.

Al preguntar sobre su experiencia al completar preguntas de examen generadas por IA, el 82 % (65 estudiantes) las calificó como "moderadamente complicadas", el 10 % (8 estudiantes) las consideró "fáciles e interesantes", mientras que el 8 % (6 estudiantes) pensó que eran "demasiado complicadas y afectaban su motivación". Estos hallazgos sugieren que, aunque en general el nivel de dificultad es adecuado, hay un grupo de estudiantes que percibe las preguntas como excesivamente complejas, lo cual podría afectar negativamente su experiencia de aprendizaje.

4. Cauteloso optimismo del profesorado durante las entrevistas. En las entrevistas, el profesorado reconoció el valor de la IA para ahorrar tiempo y aumentar la diversidad de tipos de preguntas. No obstante, también señalaron que la IA aún carece de profundidad y precisión en ciertos tipos de preguntas, como las analíticas y las abiertas. Consideran que las preguntas generadas por IA pueden utilizarse como recursos complementarios, pero deben ser revisadas y ajustadas manualmente para asegurar su coherencia con el plan de estudios.

Conclusión

Este estudio examinó las aplicaciones de la inteligencia artificial generativa (IAGen), específicamente ChatGPT y DeepSeek, en el curso Introducción a la Lingüística Inglesa, con énfasis en su rendimiento en la creación de materiales didácticos, preguntas de examen y corrección de respuestas. Los hallazgos presentados ponen de manifiesto el potencial de la IA para mejorar la eficiencia de la enseñanza y diversificar los métodos de evaluación, al tiempo que revelan diversos desafíos que requieren una atención adicional.

En primer lugar, si bien la IA puede asistir en la elaboración de materiales didácticos, sus explicaciones suelen carecer de profundidad o variedad de ejemplos. Por lo tanto, los materiales generados por IA son más adecuados como recursos complementarios, que requieren revisión o perfeccionamiento por parte del profesorado. En cuanto a las preguntas de examen, la calidad general es alta. Los resultados del cuestionario indican que el 58 % del alumnado no percibió diferencias significativas entre preguntas generadas por IA y por docentes humanos; un 19 % consideró que las preguntas generadas por IA eran más claras, mientras que un 14 % las encontró más difíciles de comprender. Las entrevistas con el profesorado sugieren que la IA puede reducir hasta en un 80 % el tiempo de diseño de exámenes, aunque sigue siendo necesaria una supervisión humana para garantizar la precisión, claridad, nivel de dificultad y cobertura del currículo, asegurando así la adecuación y equidad de las pruebas.

En segundo lugar, si bien las capacidades de corrección de la IA son prometedoras, aún presentan limitaciones. Las puntuaciones otorgadas por ChatGPT y DeepSeek muestran una alta concordancia con las calificaciones humanas (Kappa > 0,60), destacándose el resultado de DeepSeek (Kappa = 0,72). Sin embargo, la IA tiene dificultades con las preguntas subjetivas, especialmente al evaluar respuestas que implican pensamiento crítico y creativo, donde tiende a pasar por alto los aportes únicos del alumnado. Además, la IA no puede considerar factores como el contexto, la creatividad, el tono o el trasfondo cultural.

En tercer lugar, respecto a la aceptación de la IA en el ámbito educativo, el 73 % del alumnado creyó poder distinguir las preguntas generadas por IA, aunque su precisión real fue solo del 48 %, lo que sugiere que la IA puede generar preguntas muy similares a las elaboradas por humanos. En cuanto a la experiencia de aprendizaje, el 82 % consideró que las preguntas generadas por IA eran "moderadamente complicadas" y el 70 % expresó interés

en utilizar IA como herramienta de apoyo en otras asignaturas, lo que demuestra una actitud positiva hacia el uso de IA en el aprendizaje. No obstante, algunos estudiantes se muestran aún inseguros y coinciden en que la IA no puede reemplazar completamente la labor docente.

En conclusión, la IA puede ser una herramienta de apoyo poderosa, pero no puede sustituir por completo el rol del profesorado. El uso de la IA en la educación puede optimizarse aún más mediante futuras investigaciones, por ejemplo, mejorando la calidad del contenido generado por IA, optimizando la colaboración entre IA y docentes y aprovechando el potencial de la IA para el análisis de datos y retroalimentación, con recomendaciones individualizadas, y, lo más importante, asegurando que su aplicación respete principios éticos y promueva la equidad en las oportunidades educativas para todo el alumnado.

Referencias

Alers, H., Malinowska, A., Meghoe, G. y Apfel, E. (2024). Using ChatGPT-4 to grade open question exams. En *Lecture Notes in Networks and Systems* (vol. 919, pp. 1-9). Springer.

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E. y Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100177

Choi, J. H. (2024). How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics*, *180*(2), 214-233. https://doi.org/10.1628/jite-2024-0006

Chu, C.-H. y Liu, Y.-L. (2023). Augmented reality user interface design and experimental evaluation for human-robot collaborative assembly. *Journal of Manufacturing Systems*, *68*, 313-324. https://doi.org/10.1016/j.jmsy.2023.04.007

Chu, H. y Liu, Y. (2024). Research on reforming the training model of master's talents in design in ethnic regions of universities based on artificial intelligence. *En 2024 The 9th International Conference on Information and Education Innovations* (pp. 57-62). ACM. https://doi.org/10.1145/3664934.3664949

Cooper, G. (2023). Examining Science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, *32*(3), 444-452. https://doi.org/10.1007/s10956-023-10039-y

Daun, M. y Brings, J. (2023). How ChatGPT will change software engineering education. En *ITiCSE 2023: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education* (vol. 1, pp. 110-116). ACM. https://doi.org/10.1145/3587102.3588815

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D., y Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), 323-325. https://doi.org/10.1109/ICALT58122.2023.00100

Deng, R., Jiang, M., Yu, X., Lu, Y. y Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers and Education, 227*. https://doi.org/10.1016/j.compedu.2024.105224

Falchikov, N. y Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322. https://doi.org/10.3102/00346543070003287

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K. y McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment and Evaluation in Higher Education*, *49*(3), 363-375. https://doi.org/10.1080/02602938.2023.2241676

Gencer, A. y Aydin, S. (2023). Can ChatGPT pass the thoracic surgery exam? *American Journal of the Medical Sciences*, *366*(4), 291-295. https://doi.org/10.1016/j.amjms.2023.08.001

Kayaalp, M. E., Prill, R., Sezgin, E. A., Cong, T., Królikowska, A. y Hirschmann, M. T. (2025). DeepSeek versus ChatGPT: Multimodal artificial intelligence revolutionizing scientific discovery. From language editing to autonomous content generation—Redefining innovation in research and practice. *Knee Surgery, Sports Traumatology, Arthroscopy, 33*(5), 1553-1556. https://doi.org/10.1002/ksa.12628

Kung, J. E., Marshall, C., Gauthier, C., Gonzalez, T. A. y Jackson, J. B. (2023). Evaluating ChatGPT performance on the orthopaedic in-training examination. *The Journal of Bone and Joint Surgery, 8*(3), e23.00056. https://doi.org/10.2106/JBJS.23.0005

Leddo, J. (2021). Comparing the effectiveness of AI-powered educational software to human teachers. *International Journal of Social Science and Economic Research, 6*(3). https://doi.org/10.46609/IJSSER.2021.v06i03.015

Lee, G.-G. y Zhai, X. (2024). Using ChatGPT for science learning: A study on pre-service teachers' lesson planning. *IEEE Transactions on Learning Technologies*, 17, 1683-1700. https://doi.org/10.1109/TLT.2024.3401457

Leiker, D., Finnigan, S., Gyllen, A. R. y Cukurova, M. (2023). Prototyping the use of Large Language Models (LLMs) for adult learning content creation at scale. *arXiv*:2306.01815, 3-7. https://doi.org/10.48550/arXiv.2306.01815

Mabrito, M. (2025). Collaborating with generative AI in the English classroom. *International Journal of Technology, Knowledge and Society*, *21*(2), 1-23. https://doi.org/10.18848/1832-3669/CGP/v21i02/1-23

Markowitz, D. M. (2024). From complexity to clarity: How AI enhances perceptions of scientists and the public's understanding of science. *PNAS Nexus*, *3*(9). https://doi.org/10.1093/pnasnexus/pgae387

Markowitz, D. M., Hancock, J. T. y Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, *43*(1), 63-82. https://doi.org/10.1177/0261927X231200201

Mizumoto, A. y Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2). https://doi.org/10.1016/j.rmal.2023.100050

Peng, Y., Malin, B. A., Rousseau, J. F., Wang, Y., Xu, Z., Xu, X., Weng, C. y Bian, J. (2025). From GPT to DeepSeek: Significant gaps remain in realizing AI in healthcare. *Journal of Biomedical Informatics*, *163*, 104791. https://doi.org/10.1016/j.jbi.2025.104791

Pinto, G., Cardoso-Pereira, I., Monteiro, D., Lucena, D., Souza, A. y Gama, K. (2023). Large language models for education: Grading open-ended questions using ChatGPT. *arXiv*:2307.16696, 293-302. https://doi.org/10.1145/3613372.3614197

Rana, S., Sheshadri, T., Malhotra, N. y Mahabub Basha, S. (2024). Creating digital learning environments: Tools and technologies for success. En *Transdisciplinary teaching and technological integration for improved learning: Case studies and practical approaches* (pp. 1-21). IGI Global.

Sadler, P. M. y Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, *11*(1), 1-31. https://doi.org/10.1207/s15326977ea1101_1

Sun X., y Lin H. Investigación práctica sobre inteligencia artificial en la evaluación de exámenes - tomando como ejemplo las preguntas de geografía del examen de ingreso a la universidad de la provincia de Jiangsu de 2023. Geography Teaching, 2024(5): 21-23, 34.

Viera, A. J. y Garrett, J. M. (2008). Preliminary study of a school-based program to improve hypertension awareness in the community. *Family Medicine*, *40*(4), 264-270.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. y Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*(1). https://doi.org/10.1007/s40979-023-00146-z

Xu Wenbo, Zhou Xiaoping. "Exploración de la aplicación de modelos de lenguaje grandes tipo ChatGPT en la evaluación de cursos de enfermería - basado en pruebas con ChatGPT, Wenxin Yiyan y iFlytek Spark". China Medical Education Technology, 2024, 38(5): 567-571.

Yeadon, W., Agra, E., Inyang, O.-O., Mackay, P. y Mizouri, A. (2024). Evaluating AI and human authorship quality in academic writing through physics essays. *European Journal of Physics*, *45*(5). https://doi.org/10.1088/1361-6404/ad669d

Zhang, Y., Fen, B. W., Zhang, C. y Pi, S. (2024). Transforming music education through artificial intelligence: A Systematic literature review on enhancing music teaching and learning. *International Journal of Interactive Mobile Technologies*, *18*(18), 76-93. https://doi.org/10.3991/ijim.v18i18.50545

Zupanc, K. y Bosnić, Z. (2020). Improvement of automated essay grading by grouping similar graders. *Fundamenta Informaticae*, *172*(3), 239-259. https://doi.org/10.3233/FI-2020-1904