EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org
ISSN 2792-7660

Artículo Original / Original Article

Análisis comparativo de la evaluación humana y la evaluación basada en inteligencia artificial generativa de resúmenes científicos

A Comparative Analysis of Human and Generative AI-Based Evaluation of Scientific Abstracts

Aura López de Ramos[1]; Belka Bonnett-Bogallo[2]; Dimas Concepción[3]; Gustavo Quintero-Barreto[4]; Jarles Durán[5]; Nelly Meléndez[6]; Yuly Esteves[7]

[1,3,4,6] Centro de Investigación Educativa AIP (CIEDU AIP)
[2] Universidad Interamericana de Panamá (UIP)
[3,4] Universidad Tecnológica de Panamá (UTP)
[5,7] Universidad Pedagógica Experimental Libertador (UPEL)
[5,6,7] Universidad Internacional de Ciencia y Tecnología (UNICyT)
[6] Universidad Monteávila (UMA)

Email de correspondencia: *alopez@ciedupanama.org*

Cronograma editorial: *Artículo recibido 24/05/2025 Aceptado: 20/06/2025 Publicado: 01/07/2025*

Para citar este artículo utilice la siguiente referencia:
López de Ramos, A. L., Bonnett-Bogallo, B., Concepción, D., Quintero-Barreto, G., Durán, J., Meléndez, N., & Esteves, Y. (2025). Análisis comparativo de la evaluación humana y la evaluación basada en inteligencia artificial generativa de resúmenes científicos. EDUCA. Revista Internacional Para La Calidad Educativa, 5(2), 1-21. https://doi.org/10.55040/q8sgtr65

Contribución específica de los autores: Los autores han participado conjuntamente en todas las fases de la investigación.

Financiación: No existió financiación para este proyecto.

Consentimiento informado participantes del estudio: Se han solicitado los consentimientos informados de los participantes.

Conflicto de interés: Los autores no señalan ningún conflicto de interés.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org
ISSN 2792-7660

Resumen

El presente estudio analiza las diferencias en la evaluación de resúmenes enviados al II Congreso de Investigación Educativa COIE-CIEDU 2024, entre las valoraciones emitidas por dos expertos en el área con las generadas por una inteligencia artificial generativa. Se utilizó una misma rúbrica de evaluación, aplicando pruebas de diferencia de medias a fin de determinar la existencia de discrepancias significativas. Los resultados muestran que, si bien no se hallaron diferencias significativas entre los expertos humanos, sí se identificaron discrepancias estadísticamente significativas entre las evaluaciones humanas y las de la inteligencia artificial generativa ($p < 0,05$). Este hallazgo evidencia que, aunque el juicio humano mantiene una consistencia metodológica, la inteligencia artificial generativa no logra aún emular los estándares de calidad aplicados por revisores expertos. Se concluye que la inteligencia artificial generativa, aunque útil como herramienta de apoyo en tareas técnicas o administrativas del proceso de revisión, no está aún preparada para desempeñar de forma autónoma funciones de arbitraje académico. Se recomienda su implementación como complemento, bajo protocolos de supervisión humana y con validación continua de su desempeño, a fin de garantizar la equidad, la rigurosidad y la integridad en la evaluación de contenidos científicos.

Palabras clave: análisis comparativo, inteligencia artificial, resumen, evaluación, investigación educativa.

Abstract

This study analyzes the differences in the evaluation of abstracts submitted to the II Educational Research Congress COIE-CIEDU 2024, comparing the assessments made by two subject-matter experts with those generated by a generative artificial intelligence system. A standardized evaluation rubric was used, and mean difference tests were applied to determine the presence of statistically significant discrepancies. The results indicate that, while no significant differences were found between the human experts, statistically significant discrepancies were identified between the human evaluations and those generated by the generative artificial intelligence system ($p < 0.05$). This finding demonstrates that, although human judgment maintains methodological consistency, generative artificial intelligence is not yet capable of replicating the academic quality standards applied by expert reviewers. It is concluded that, although generative artificial intelligence may serve as a valuable support tool for technical or administrative tasks within the review process, it is not ready to autonomously perform academic peer-review functions. Its implementation is recommended as a complementary resource, under clear supervision protocols and continuous performance validation, in order to ensure fairness, rigor, and integrity in the evaluation of scientific content.

Keywords: comparative analysis, artificial intelligence, abstract, evaluation, educational research.

.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org
ISSN 2792-7660

Introducción

La revisión por pares es un componente esencial en la validación de publicaciones académicas, asegurando la calidad y rigurosidad de los estudios publicados (Harris & Davison, 2020). Tradicionalmente, este proceso ha sido realizado por expertos humanos, pero con la creciente implementación de herramientas de inteligencia artificial (IA) en diversas áreas, ha surgido el interés por evaluar su desempeño en la revisión de resúmenes académicos (Jiang, 2024). Estudios previos han sugerido que la IA puede ser efectiva en tareas de preselección y detección de errores formales, aunque aún existe incertidumbre sobre su capacidad para evaluar aspectos conceptuales y metodológicos (Nematov, 2025; Acosta Camino & Andrade Clavijo, 2024; Mondal & Mondal, 2024; Mostafapour et al., 2024; Meléndez et al., 2023; Cheng et al., 2023; Liu & Shah, 2023).

Este estudio tiene como objetivo analizar las diferencias en la evaluación de resúmenes enviados al II Congreso de Investigación Educativa COIE-CIEDU 2024, comparando las valoraciones emitidas por dos expertos en el área con las generadas por una Inteligencia Artificial (IA).

Descripción y definición del proceso de arbitraje (revisión por pares)

En las publicaciones académicas se utiliza el proceso de arbitraje por pares como componente esencial para garantizar la calidad de los productos que se comparten al público. Por lo general son expertos en el campo quienes, con base en las normas editoriales de las revistas, registran los resultados del análisis en el baremo provisto por los editores. Estos artículos se entregan de forma anónima al evaluador y el investigador desconoce quién es su evaluador, todo el tiempo es la revista el intermediario entre ambas partes.

Entre las ventajas que pueden enumerarse de este proceso se indica que la revisión por pares mejora la calidad de las publicaciones académicas al asegurar que los artículos cumplan con altos estándares y filtran trabajos de baja calidad o con afirmaciones no justificadas (Harris & Davison, 2020; Riding, 2022). Adicionalmente, la diversidad de perspectivas de evaluadores ayuda a eliminar sesgos personales y permite identificar y rechazar investigaciones duplicadas o plagiadas, protegiendo de esta manera la integridad científica (Faintuch & Faintuch, 2022). Adicionalmente, ser invitado a revisar es un reconocimiento profesional que puede abrir las puertas para las actividades editoriales (Harris & Davison, 2020; Riding, 2022).

No obstante, aún persisten las preocupaciones sobre la imparcialidad, ya que los revisores podrían tener conflictos de intereses o sesgos inconscientes, así como alta carga de trabajo que afecte las revisiones (González et al., 2022); también la lentitud en el proceso de revisión en algunos casos tiende a retrasar la publicación de nuevos hallazgos (Kelly et al., 2014).

Autores como Heesen & Bright (2021) incluso llegan a sugerir que se debería abolir la revisión por pares a la publicación, porque esta práctica afecta el comportamiento de los investigadores y mantiene conceptualizaciones paradigmáticas que aumentan el dominio del *status quo*.

En este estudio se considera que la revisión por pares es un pilar fundamental en la validación de la investigación académica, pues garantiza la calidad, el rigor y la credibilidad de las publicaciones. Si bien, este proceso ha sido tradicionalmente llevado a cabo por expertos humanos, la creciente integración de herramientas de inteligencia artificial (IA) ha suscitado interés en su potencial para complementar o incluso transformar la revisión académica. Por tal motivo, se analizan las diferencias en la evaluación de resúmenes enviados al COIE-CIEDU 2024 entre las valoraciones emitidas por dos expertos en el área con las generadas por una IA con el fin de determinar su potencial para contribuir a la revisión por pares y evaluación de textos.

Inteligencia artificial en la revisión académica

La revisión de literatura destaca hallazgos que contribuyen al presente estudio. Entre éstas, cabe resaltar la investigación realizada por Meléndez et al. (2023) la cual se centró en la aplicación de metodologías de Modelos de Lenguaje de Gran Escala (LLM: Large Language Models) con el propósito de identificar el potencial de estas herramientas para la evaluación de ensayos redactados por estudiantes. El estudio fue experimental, usando “Prompting Engineering” o "Ingeniería de Generación de Instrucciones" (IGI), que permitió configurar la IA para la ejecución de tareas académicas específicas.

La metodología de evaluación implicó la exposición a cada LLM de dos configuraciones distintas de instrucciones IGI. Inicialmente, se empleó una plantilla de instrucciones básicas que contenía un conjunto de directrices generales y el texto del ensayo a evaluar, pero sin

proporcionar mayores referencias. Posteriormente, se repitió la prueba usando una plantilla más elaborada que contenía el texto fuente con instrucciones para la LLM (rúbrica) y el ensayo.

Los resultados demostraron que ambas estrategias fueron eficaces para evaluar ensayos académicos. No obstante, se observó una mejora significativa en la precisión de la retroalimentación cuando se le suministraba una fuente textual adecuadamente construida y con instrucciones detalladas. En estos casos, la evaluación reflejó una mayor exigencia en cuanto a la calidad del ensayo y una objetividad superior en los hallazgos.

Esta investigación se relaciona con el presente artículo porque demuestra la posibilidad de evaluar eficazmente textos mediante el uso de una rúbrica, instrumento que también se ha empleado en la evaluación de resúmenes académicos presentados para el congreso tomado como muestra.

Otro aspecto estudiado es cómo la IA puede mejorar la primera revisión de revistas académicas dado su potencial para evaluar la novedad de los manuscritos, estandarizar el formato, realizar recomendaciones de revisores expertos y generar opiniones objetivas sobre el texto. Todo esto abordado en una investigación realizada por Jiang (2024) quien concluyó que existen evidencias suficientes para confiar en el potencial de la IA en la revisión de artículos y la selección de árbitros.

En indagación de fuentes relacionadas con el uso de IA para la selección de evaluadores se encontró el estudio de Farber (2024), quien en un análisis de métodos mixtos propuso la selección de revisores con inteligencia artificial en distintas disciplinas académicas, reduciendo en un 73% el tiempo de selección de revisores, pero en el campo de las ciencias sociales e interdisciplinarios la IA sólo mostró un 35% de precisión. Los editores expresaron también su preocupación por el sesgo algorítmico, las disparidades geográficas e institucionales y los casos donde la IA sugiere revisores ficticios (un 10% de los casos), lo que destaca la necesidad de supervisión humana y desarrollar sistemas personalizados para diferentes procesos propios de la evaluación por pares.

En cuanto al uso de IA para apoyar la revisión por pares, Kousha & Thelwall (2024) encontraron que la inteligencia artificial además de ser útil para conseguir revisores, también puede ayudar al control inicial de manuscritos enviados a revistas académicas, apoyar la

revisión de pares y la publicación académica, pero para la fecha del estudio (2023) no parecía capaz de sustituir la revisión por pares en manuscritos enviados a revistas académicas.

Entre las referencias analizadas hay un caso donde se utilizó la herramienta tecnológica de IA de software libre "ASReview" en la revisión de títulos y resúmenes académicos (van Dijk et al., 2023). Primero, se entrenó el algoritmo con varios artículos etiquetados antes de la revisión. Luego, utilizando un algoritmo de investigador, la herramienta de IA propuso el artículo con la mayor probabilidad de ser relevante. ASReview determina el nivel de relevancia de cada artículo basándose en la probabilidad que le asigna el modelo de aprendizaje automático entrenado. Dos elementos clave que influyen en esta probabilidad son:

- Características del Texto (Text Features): Durante la fase de entrenamiento, el modelo aprende a identificar patrones lingüísticos y características textuales que son predictivas de la relevancia.
- Incertidumbre del Modelo (Model Uncertainty): Como se mencionó en los pasos del algoritmo, ASReview prioriza los documentos donde el modelo tiene la mayor incertidumbre sobre su clasificación.

Luego, el software revisor decidió la relevancia de cada artículo propuesto. Este proceso continuó hasta que se alcanzó el criterio de parada. Todos los artículos etiquetados como relevantes por el revisor se examinaron en texto completo por un humano.

Para asegurar la calidad metodológica al utilizar IA en revisiones sistemáticas se incluyeron con "ASReview": la elección de si utiliza o no IA, la verificación del acuerdo entre revisores, elegir un criterio de parada y la calidad del informe. El uso de la herramienta en la revisión resultó en un gran ahorro de tiempo: solo el 23% de los artículos fueron evaluados por el revisor humano. Por lo cual, los investigadores concluyen que la herramienta de IA es una innovación prometedora para la práctica actual de revisión sistemática, siempre que se utilice adecuadamente y se pueda asegurar la calidad metodológica, por lo que es fundamental tomar en cuenta que:

- La fase de entrenamiento inicial con ASReview requiere tiempo y esfuerzo del revisor. La calidad de las etiquetas iniciales es crucial para el rendimiento del modelo.
- El ahorro de tiempo es más significativo cuanto mayor sea el volumen de literatura a revisar y cuanto más preciso se vuelva el modelo. Para conjuntos de datos muy

pequeños, la sobrecarga inicial de entrenar el modelo podría no justificar el ahorro de tiempo.

- Se requiere cierta curva de aprendizaje para familiarizarse con el software y entender cómo interactúa el algoritmo.

Los estudios mencionados solo son una muestra del creciente interés acerca de la eficacia en escritura académica, la cual puede ser abordada desde perspectivas multifacéticas, pero que puede proveer mejoras significativas en términos de eficiencia, precisión y objetividad (Salman et al., 2025); por lo que las herramientas de IA son prometedoras si se utilizan las instrucciones y entrenamiento adecuados (Kharipova et al., 2024).

Metodología

Diseño del estudio

Esta investigación se enmarca en un diseño descriptivo-comparativo, que permite caracterizar y contrastar las puntuaciones de tres evaluadores (dos humanos y una máquina) sobre un mismo conjunto de unidades de análisis. Este diseño metodológico permite identificar estadísticamente la magnitud y significancia de las diferencias entre los datos de análisis y se justifica dado que:

- El estudio busca mostrar y caracterizar los puntajes de los resúmenes por cada una de las tres fuentes de evaluación (humano 1, humano 2 e IA). Estos puntajes varían entre 0 y 15 puntos.
- El objetivo central es contrastar los puntajes asignados a los resúmenes entre los diferentes grupos de evaluadores (experto 1 vs. experto 2, experto 1 vs. IA y por último experto 2 vs. IA) . Esto implica realizar el cálculo de las diferencias entre los puntajes de cada par de evaluadores antes señaladas, con la finalidad de calcular las estadísticas descriptivas como medidas de tendencia central y de dispersión para realizar el contraste de hipótesis que permitirá analizar si existen diferencias significativas en las puntuaciones asignadas por cada uno de los expertos.

Entre las fortalezas de este diseño se encuentran:

- Aleatorización de la unidad de análisis: La selección aleatoria de los 50 resúmenes asegura que la muestra sea representativa del conjunto total de resúmenes enviados al congreso, lo que aumenta la validez externa de los resultados.

- Evaluación independiente: Realizar las evaluaciones de forma independiente con la misma rúbrica garantiza que cada evaluador (experto o IA) aplique los mismos criterios sin influencia de los demás, lo que fortalece la validez interna.
- Aseguramiento de la confidencialidad y equidad: Estas medidas éticas son cruciales para minimizar sesgos y garantizar la integridad del proceso de evaluación.

Unidades de análisis

La recolección de datos para este estudio se centró en tres fuentes de evaluación (dos humanas, una artificial) que analizaron un mismo corpus de resúmenes de congreso. Primeramente, se recopilaron las puntuaciones asignadas por los dos evaluadores humanos (docentes). Esta recopilación se realizó una vez finalizado el evento y se formalizó mediante la obtención de un consentimiento informado por escrito, garantizando el uso ético de los datos. Todos los resúmenes fueron anonimizados (excluyendo los autores, instituciones de adscripción, propiedades del documento y datos de los pares que los evaluaron). A continuación, y tras la publicación oficial de los resúmenes en la página web del congreso, se procedió a la evaluación automatizada de los mismos mediante un sistema de inteligencia artificial. En total se incluyeron 50 resúmenes enviados al II Congreso de Investigación Educativa COIE-CIEDU 2024.

Variables

Se recabaron las valoraciones cuantitativas asignadas por:

- Experto evaluador 1
- Experto evaluador 2
- Inteligencia Artificial

Medición y seguimiento

La revisión de los resúmenes se realizó de manera independiente con la misma rúbrica de evaluación establecida por el congreso (Tabla 1). Es decir, que la IA se entrenó con la misma rúbrica que usaron los expertos 1 y 2. También, la IA evaluó las mismas versiones de los resúmenes que usaron los expertos 1 y 2. Se aseguró la confidencialidad de los evaluadores y la equidad en el proceso.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org
ISSN 2792-7660

Criterio	3	2	1	0
Justificación (Relevancia científica y social).	Alta necesidad de llevar a cabo el estudio es establecida en cuanto a su relevancia científica y social.	Cierta necesidad de llevar a cabo el estudio es establecida en cuanto a su relevancia científica y/o social	Baja necesidad de llevar a cabo el estudio es establecida en cuanto a su relevancia científica o social.	No se establece la necesidad de llevar a cabo el estudio.
Pregunta de Investigación /Objetivos	No se presenta el propósito del estudio o pregunta de investigación.	El propósito del estudio es presentado y preguntas, objetivos o hipótesis son proporcionadas con inconsistencias inapropiadas.	El propósito del estudio está presentado y preguntas, objetivos o hipótesis apropiadas son proporcionadas con cierta claridad.	El propósito del estudio está claramente presentado y preguntas, objetivos o hipótesis apropiadas son proporcionadas.
Análisis y recolección de data	Los métodos de recolección y análisis de datos son apropiados para este estudio.	Se presentan sólo los métodos de recolección o de análisis de datos para este estudio, y son apropiados.	Se presentan métodos de recolección de datos para este estudio con inconsistencias inapropiadas o falta de claridad.	No se presentan métodos de recolección y análisis de datos, o no son los apropiados.
Resultados	Datos apropiados de resultados o evidencia de sustento son presentados clara y lógicamente.	Datos apropiados de resultados o evidencia de sustento son presentados con cierta claridad y lógica.	Datos apropiados de resultados o evidencia de sustento son presentados con inconsistencias inapropiadas o falta de claridad.	No se presentan resultados, o no son los apropiados.
Discusión y conclusiones	Los datos son interpretados correctamente, y las conclusiones son apropiadas y presentadas con claridad.	Los datos son interpretados correctamente, y las conclusiones apropiadas	Los datos son interpretados con inconsistencias o las conclusiones son poco claras.	Los datos son interpretados incorrectamente o las conclusiones son inapropiadas, o no se presentan interpretación o conclusiones.
Puntaje total máximo			15	

Tabla 1. Rúbrica de evaluación utilizada en el congreso

Se seleccionó ChatGPT 4o para la evaluación de los artículos. Las razones para esta selección fueron las siguientes: tiene consistencia en la evaluación ya que aplica criterios de

forma uniforme; facilita la rápida identificación de objetivos, metodología y resultados claves de los resúmenes científicos; permite evaluar un alto número de resúmenes en corto tiempo.

La técnica empleada en la redacción de los prompts correspondió a CLEAR: C – Context (Contexto); L – Length (Longitud); E – Explicitness (Claridad o Explicitud); A – Audience (Audiencia); R – Response Format (Formato de respuesta). Los prompts fueron claros, legibles, enfocados, adaptados al contexto de la evaluación científica y revisables. Se redactó una estructura lógica orientada a obtener respuestas consistentes y rigurosas, propias de tareas académicas con criterios definidos de calidad.

Análisis de datos

La técnica de análisis estadístico se sustenta en los contrastes de hipótesis por medio de intervalos de confianza, como prueba de diferencia entre los puntajes que se muestran a continuación:

- Experto 1 vs. Experto 2
- Experto 1 vs. IA
- Experto 2 vs. IA

Para construir el estadístico de prueba (intervalos de confianza), se utilizó la distribución Z con un nivel de significancia del 0,05, siguiendo el Teorema Central del Límite para garantizar la normalidad de los datos; esto, porque, para poder formalizar los procedimientos estadísticos es necesario cumplir con el supuesto de normalidad de los datos, es allí donde este teorema se convierte en parte determinante en esta investigación. Como lo plantean Anderson et al. (2008) "cuando se seleccionan muestras aleatorias simples de tamaño *n* de una población, la distribución muestral de la media puede aproximarse mediante una distribución normal a medida que el tamaño de la muestra se hace grande" (p. 272), esto para garantizar que los resultados pueden ser generalizados a toda la población (hacer inferencia). Aplicando este teorema, no importa la distribución de probabilidad original de los datos, pues, si la población o muestra es lo suficientemente grande, entonces tendrán una aproximación a la distribución de probabilidad normal estándar.

Teniendo resuelta la normalización de los datos, se continuó con el análisis estadístico. Se realizaron pruebas de diferencias sobre los puntajes asignados a cada resumen tanto para ambos expertos como para la IA, es decir, se comparó los puntajes asignados por el primer

experto a cada resumen vs el puntaje asignado por el segundo experto a los resúmenes ya evaluados por el primer experto; luego se comparó el puntaje de los resúmenes por parte del primer experto vs la puntuación asignada por la IA a esos mismos resúmenes y por último se compara el puntaje asignado por el segundo experto vs el puntaje de la IA.

El tratamiento estadístico aplicado para demostrar si existe diferencia significativa entre el puntaje asignado a cada experto y la IA se sustenta en el contraste de hipótesis sobre las diferencias en promedio, una vez superado el supuesto de normalidad que exige la prueba de diferencia de medias, por lo que se procede a realizar los cálculos de las estadísticas de tendencia central y de dispersión como insumos indispensables para dicha prueba. Por tanto, se procede a realizar los procedimientos que se ameritan.

Procedimiento descriptivo: se describen las diferencias en promedio entre los puntajes asignados por los expertos a través de las estadísticas básicas de tendencia central y dispersión. Los datos se presentan en la Tabla 2.

Estadístico	Dif. Experto 1-2	Dif. Experto 1-IA	Dif. Experto 2-IA
Media	0,12	-2,48	-2.6
Varianza	12,9648	16,2955	15,1
Observaciones	50	50	50

Tabla 2. Estadísticas básicas de las diferencias entre experto 1, experto 2 e IA. Año 2024

En la Tabla 2 se observa una diferencia promedio de 0,12 puntos entre el experto 1 vs experto 2, indicando que existe una diferencia mínima entre ambos evaluadores, siendo esta la diferencia promedio la más baja de la investigación, es decir que el experto 1 y el experto 2 tienden a ser consistentes en la evaluación del resumen (entre humanos). También se observa una diferencia promedio de -2,48 puntos entre el experto 1 y la IA, lo que indica que, en promedio, la IA asignó puntuaciones 2,48 puntos superiores a las del experto 1. Por último, se observa una diferencia promedio de -2,6 puntos entre el puntaje del experto 2 vs el puntaje de la IA, indicando nuevamente que la IA asignó puntajes superiores a los propuestos por el experto 2.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org
ISSN 2792-7660

En conclusión, se sospecha que la IA asigna puntuaciones superiores a los puntajes asignados por los expertos 1 y 2. Esta situación es analizada en el procedimiento estadístico de contraste de hipótesis como método de comparación que se muestra a continuación:

Procedimiento comparativo: se usa el método estadístico de contraste de hipótesis para detectar diferencias significativas entre los evaluadores, por tanto, se describen las hipótesis que se desean contrastar. La hipótesis nula (H_0) indica que no existe diferencia significativa entre el puntaje asignado por el experto 1 vs. experto 2; el experto 1 vs. IA; y por último experto 2 vs. IA. Esta hipótesis (H_0) se contrasta con la hipótesis alternativa (H_1) la cual sostiene que si existe diferencia significativa entre el puntaje asignado por el experto 1 vs. experto 2; el experto 1 vs. IA; y por último experto 2 vs. IA. Es importante señalar que la naturaleza de la investigación se centra en las diferencias entre los puntajes asignados por los tres evaluadores (experto 1, experto 2 y la IA) sobre los resúmenes enviados al II Congreso de Investigación Educativa COIE-CIEDU 2024 y no sobre el promedio total de cada evaluador ya que se estaría distorsionando la intención investigativa.

Cada procedimiento de análisis estadístico posee sus estadísticos de prueba específicos según la naturaleza de la investigación, en este caso de usó el siguiente intervalo de confianza de la Distribución Normal Estándar con un nivel de significancia del 5%:

Estadístico de prueba

$$\underline{X}_{d(i-j)} \pm Z_{\left(\frac{\alpha}{2}\right)}\sqrt{\frac{s^2 d(i-j)}{n}} \leq \mu_{(i-j)} \qquad \forall\, i \neq j; i = 1{,}2;\ j = 2, IA \ (1)$$

El estadístico para comprobar según tabla de la Distribución Normal Estándar con un nivel de significación de 0,05 para dos colas es de:

$$Z_{0{,}025} = 1{,}96 \ \ (2)$$

Para realizar el contraste de hipótesis se requiere de las estadísticas descriptivas tales como las medidas de tendencia central y de dispersión, para ello se usaron la media y la varianza mostradas en la Tabla 2. Se agregan los estadísticos z, grados de libertad (gl), prueba de significación (Sig) también abreviado P-value y los límites de los intervalos de confianza (Límite infe. y Límite supr.). Los resultados del contraste de hipótesis se muestran en la Tabla 3.

Prueba de muestra única
Valor de prueba = 0

Contraste	Media	Varianza	n	z	gl	Sig. (bilateral)	95% de intervalo de confianza de la diferencia	
							Límite infe.	Límite supr.
Dif. Experto 1-2	0,12	12,965	50	0,236	49	0,815	-0,878	1,111
Dif. Experto 1-IA	-2,48	16,296	50	-4,344	49	0,000*	-3,599	-1,361
Dif. Experto 2-IA	-2.6	15,100	50	-4,737	49	0,000*	-3,677	-1,522

Tabla 3. Contraste de hipótesis por medio de intervalos de confianza para las diferencias en promedio. Año 2024

Nota: * son las pruebas significativas al 0,05

En la Tabla 3 se aprecia que las diferencias entre los puntajes asignados por el experto 1 y el experto 2 no rechazan la Hipótesis nula H_O, es decir, no existe diferencia significativa entre los puntajes de los expertos 1 y 2 asignados a los resúmenes enviados al II Congreso de Investigación Educativa COIE-CIEDU 2024 con un nivel de confianza del 95%; sin embargo, existe diferencia significativa entre los puntajes asignados por el experto 1 y la IA, de la misma manera ocurre entre el experto 2 y la IA al mismo nivel de confianza del 95%. Como puede observarse, los puntajes asignados por la IA son estadísticamente diferentes e incluso superiores a los asignados por los expertos 1 y 2.

Resultados

Los análisis de diferencia de medias mostraron que:

- No hubo diferencias significativas entre los puntajes asignados a los resúmenes por parte de los expertos 1 y 2 (humanos) con un nivel de significancia del 5% ($p > 0.05$).
- Se encontraron diferencias significativas entre los puntajes asignados a los resúmenes entre la IA y la de ambos expertos 1 y 2 con un nivel de significancia del 5% ($p < 0.05$), lo que indica discrepancias en los criterios de valoración utilizados por la IA en comparación con los revisores humanos.

EDUCA Revista Internacional para la calidad educativa
EDUCA International Journal for Educational Quality
Publisher: Campus EDUCA SPORTIS. A Coruña. Spain | E-mail: info@plataformaeduca.org | Web: www.revistaeduca.org
ISSN 2792-7660

El hecho de que no haya diferencias significativas entre los puntajes de los expertos humanos sugiere cierta consistencia en su evaluación, lo cual es positivo. Sin embargo, las diferencias significativas entre la IA y ambos expertos apuntan a una discrepancia fundamental en cómo la IA está evaluando los resúmenes en comparación con el juicio humano. Esto podría indicar que el algoritmo de IA, en su estado actual, no está alineado con los criterios de valoración que los expertos humanos consideran importantes. Esta diferencia sugiere una necesidad de revisar, refinar y recalibrar el modelo de IA.

Con el fin de facilitar la comprensión de los resultados obtenidos en esta investigación se presenta la figura 1 que se muestra de manera didáctica el concepto asociado al contraste de hipótesis de diferencia de medias, en este caso en particular se observa que el contraste entre el experto 1 y el experto 2 contienen el valor 0 indicando que no existe diferencia significativa entre los puntajes asignados a los resúmenes; mientras que, los contrastes entre los expertos 1 y 2 vs la IA se alejan del valor central de manera considerable.

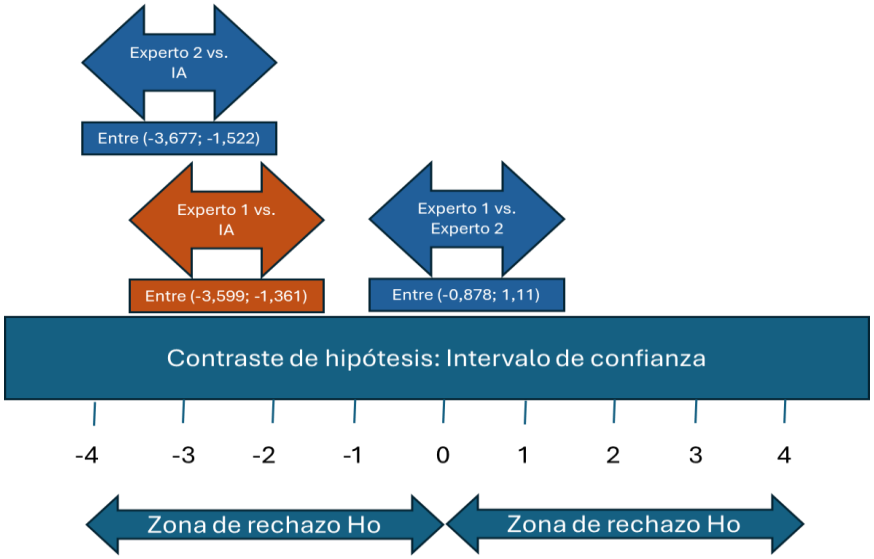

Figura 1. Concepto de contraste de hipótesis de diferencia de medias

También puede ser beneficioso incorporar explícitamente en el modelo de IA los criterios de evaluación definidos en la rúbrica, quizás a través de la ingeniería de características o mediante el uso de técnicas de PNL más avanzadas que puedan comprender y aplicar estos criterios de manera más similar a los humanos. Entender por qué la IA asigna ciertas valoraciones (qué características del texto está priorizando) podría proporcionar información

valiosa para identificar las áreas de desacuerdo con los humanos y guiar el proceso de refinamiento.

El hallazgo de diferencias significativas plantea serias cuestiones sobre la viabilidad y la implementación de la IA como un sustituto directo o un filtro principal para la evaluación humana en el arbitraje de resúmenes científicos, al menos con el modelo actual. Si la IA consistentemente valora los resúmenes de manera diferente a los expertos humanos:

- Podría llevar a decisiones subóptimas, ya que resúmenes que la IA considera de alta calidad podrían ser rechazados por los humanos, y viceversa, lo que podría afectar la calidad general y la equidad del proceso de selección para el congreso.
- Podría erosionar la confianza en el sistema de revisión, ya que los investigadores podrían percibir que la IA está utilizando criterios de evaluación diferentes y potencialmente menos válidos que los expertos humanos, creando una resistencia a la adopción de sistemas basados en IA en el proceso de revisión.

En lugar de una evaluación autónoma, la IA podría ser más útil como una herramienta de apoyo para los revisores humanos, por ejemplo, ayudando en la detección de errores formales o en la identificación de posibles conflictos de interés, pero manteniendo la evaluación del contenido y la calidad en manos de los expertos humanos.

En resumen, la discrepancia significativa entre la IA y los evaluadores humanos subraya la complejidad de replicar el juicio humano en la evaluación científica y destaca la necesidad de un desarrollo y una validación cuidadosos antes de confiar plenamente en la IA para tareas críticas como el arbitraje de resúmenes.

Discusión

Los resultados sugieren que la IA genera valoraciones significativamente distintas a las de los expertos humanos, lo que podría atribuirse a diferencias en la interpretación de los criterios de evaluación o a limitaciones en la comprensión contextual de los textos. Esto coincide con estudios previos que han encontrado que la IA es eficaz en evaluaciones técnicas, pero presenta dificultades en aspectos críticos y argumentativos (Jiang, 2024; Meléndez et al., 2023). Esta brecha también es coherente con hallazgos previos, como los de Hsu (2023), quien advierte que, aunque la IA puede asistir en tareas de redacción y evaluación, su uso autónomo compromete los estándares académicos y requiere supervisión humana rigurosa.

Este estudio difiere de resultados previos que encuentran un alto nivel de acuerdo con el revisar resúmenes, aunque en las evaluaciones de artículos completos demuestran un bajo nivel de acuerdo entre IA y revisores humanos (Shcherbiak et al., 2024). No obstante, es factible una colaboración entre humanos e IA para evaluar la novedad de los manuscritos, estandarización de los formatos y hacer recomendaciones a revisores expertos (Jiang, 2024; Mondal et al., 2023).

En términos prácticos, este hallazgo plantea retos sustantivos para la implementación de la IA como herramienta de arbitraje académico. Si se emplea sin calibración, puede generar decisiones de aceptación o rechazo inconsistentes con los estándares humanos, afectando la equidad del proceso de selección. No obstante, en línea con lo señalado por Nguyen et al. (2024), se observa que cuando los usuarios interactúan con la IA de forma iterativa y crítica, los resultados mejoran. Esto sugiere que su uso más prometedor está en la modalidad colaborativa, como apoyo a revisores humanos, especialmente en tareas preliminares como la detección de errores formales o el filtrado inicial de manuscritos.

Un aspecto para considerar en futuras investigaciones es la comparación entre los resultados de revisores humanos y la IA para discernir entre autores humanos y texto generado mediante inteligencia artificial generativa, dado que investigaciones demuestran que los revisores humanos tienen mayores dificultades para identificar correctamente la autoría de textos generados por inteligencia artificial (Shcherbiak et al., 2024; Yeadon et al., 2024; Silva et al., 2024; Dergaa et al., 2023; Liu et al., 2024).

Además, el impacto práctico de los hallazgos de este estudio se vincula con las posibilidades de integrar la IA en los sistemas de gestión editorial académica. Vuong et al. (2023) argumentan que la era de la IA en la publicación científica ya ha comenzado y que se requiere un enfoque proactivo para su adopción responsable. Por tanto, los resultados aquí expuestos pueden contribuir a desarrollar políticas de uso que equilibren eficiencia, transparencia y equidad.

Finalmente, en coherencia con los aportes de Koga (2023), debe destacarse que el uso práctico de IA exige no solo nuevas competencias de evaluación y edición por parte de los revisores, sino también la formulación de directrices institucionales que garanticen la trazabilidad y la responsabilidad en su aplicación. De este modo, se fortalece la confianza en el

ecosistema editorial y se prepara el terreno para futuras sinergias entre tecnologías emergentes y la comunidad científica.

Conclusiones

Este estudio confirma que, aunque la IA puede ser una herramienta valiosa en la evaluación académica, su aplicación en revisión por pares aún presenta desafíos significativos. Es recomendable desarrollar algoritmos más avanzados que mejoren su capacidad de comprensión de textos académicos y considerar su uso como complemento, más que sustituto, del arbitraje humano.

Si bien la IA podría complementar la revisión por pares agilizando el proceso y reduciendo sesgos humanos, su uso como evaluador primario requiere una calibración más precisa para alinearse con los criterios académicos establecidos. Por lo que la investigación futura debe seguir explorando estas sinergias, asegurando que la integridad y calidad de la revisión académica se mantengan en un entorno en constante evolución. A pesar de que algunas investigaciones recientes sugieren que la IA puede ser efectiva en la preselección de manuscritos y en la detección de errores formales, en esta investigación se encontró que su capacidad para evaluar la profundidad conceptual y metodológica de los trabajos sigue siendo incierta.

Finalmente, se reconoce que el rol de la IA en la evaluación académica es dinámico y seguirá evolucionando. Por ello, futuras investigaciones deberán profundizar en estrategias de entrenamiento supervisado, validación cruzada y colaboración humano-IA, con el objetivo de lograr un equilibrio que mantenga la integridad, la equidad y la calidad del arbitraje académico en un entorno cada vez más digitalizado.

Limitaciones del estudio y futuras líneas de investigación

La discrepancia significativa encontrada entre los evaluadores humanos y ChatGPT indican que existe la necesidad de revisar, refinar y recalibrar el modelo de IA utilizado. Futuros estudios podrían explorar la calibración de IA con aprendizaje supervisado y su aplicación en otros tipos de revisión académica.

En el estudio se utilizó un único modelo de IA (ChatGPT 4o) y una única estrategia de prompt design (CLEAR), sin explorar cómo variaciones en la formulación de instrucciones o

el uso de otros modelos de lenguaje (LLM) podrían modificar los resultados. Estudios futuros podrían comparar múltiples arquitecturas de IA y estrategias de interacción para identificar las condiciones que optimicen la alineación entre el juicio humano y el de la máquina.

Además, este estudio no abordó la percepción de los evaluadores humanos respecto al uso de IA como herramienta de apoyo ni analizó la trazabilidad de los criterios utilizados por la IA para emitir sus juicios. En este sentido, futuras investigaciones podrían combinar metodologías cuantitativas y cualitativas, integrando entrevistas, análisis de trazabilidad de tokens y estudios de caso para comprender mejor los factores que influyen en la discrepancia valorativa.

Aunque los resultados de este estudio ofrecen un primer diagnóstico útil, es necesario seguir investigando con diseños más robustos y ampliados que contribuyan al desarrollo de marcos metodológicos y normativos para la integración ética, eficiente y transparente de la inteligencia artificial en los procesos de evaluación científica.

Consideraciones éticas

Se respetaron principios de confidencialidad y transparencia, asegurando el anonimato de los evaluadores y la utilización de la IA bajo criterios de equidad y no discriminación. Además, se solicitó el consentimiento informado a los evaluadores para utilizar sus valoraciones en este estudio.

Referencias

Acosta Camino, D. F., & Andrade Clavijo, B. P. (2024). La inteligencia artificial en la investigación y redacción de textos académicos. Espíritu Emprendedor TES, 8(1), 19–34. https://doi.org/10.33970/eetes.v8.n1.2024.369

Anderson, D., Sweeney, D., & Williams, T. (2008). Estadística para administración y economía. 10a edición. Cengage Learning Editores, S.A.

Cheng, S. L., Tsai, S. J., Bai, Y. M., Ko, C. H., Hsu, C. W., Yang, F. C., ... & Su, K. P. (2023). Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. Journal of Medical Internet Research, 25(1), e51229. https://doi.org/10.2196/51229

Dergaa, I., Chamari, K., Żmijewski, P., & Ben Saad, H. (2023). From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. Biology of sport, 40(2), 615-622. https://doi.org/10.5114/biolsport.2023.125623

Faintuch, J., & Faintuch, S. (2022). Integrity of Scientific Research: Fraud, Misconduct and Fake News in the Academic, Medical and Social Environment. Springer Nature. https://doi.org/10.1007/978-3-030-99680-2

Farber, S. (2024). Enhancing peer review efficiency: A mixed-methods analysis of artificial intelligence-assisted reviewer selection across academic disciplines. Learned Publishing, 37(4), Article e1638. https://doi.org/10.1002/leap.1638

González, P., Wilson, G., & Purvis, A. (2022). Peer review in academic publishing: Challenges in achieving the gold standard. Journal of University Teaching and Learning Practice, 19(5). https://doi.org/10.53761/1.19.5.1

Harris, R. W., & Davison, R. M. (2020). Peer review: Academia's most important but least understood task. The Electronic Journal of Information Systems in Developing Countries, 86(6), isd212150. https://doi.org/10.1002/isd2.12150

Heesen, R., & Bright, L. K. (2021). Is Peer Review a Good Idea? The British Journal for the Philosophy of Science, 72(3), 635-663. https://doi.org/10.1093/bjps/axz029

Jiang, Y. (2024). An Exploration of AI Aid to the First Review of Academic Journals. Journal of New Media and Economics, 1(3), 97-100. https://doi.org/10.62517/jnme.202410317

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer Review in Scientific Publications: Benefits, Critiques, & A Survival Guide. EJIFCC, 25(3), 227 - 243. https://pmc.ncbi.nlm.nih.gov/articles/PMC4975196/

Kharipova, R., Khaydarov, I., Akramova, S., Lutfullaeva, D., Saidov, S., Erkinov, A., Azizkhonova, S., & Erkinova, N. (2024). The Role of Artificial Intelligence Technologies in Evaluating the Veracity of Scientific Research. Journal of Internet Services and Information Security, 14(4), 554-568. https://doi.org/10.58346/JISIS.2024.I4.035

Koga, S. (2023). The Integration of Large Language Models Such as ChatGPT in Scientific Writing: Harnessing Potential and Addressing Pitfalls. Korean journal of radiology, 24(9), 924-925. https://doi.org/10.3348/kjr.2023.0738

Kousha, K., & Thelwall, M. (2024). Artificial intelligence to support publishing and peer review: A summary and review. Learned Publishing, 37(1), 4-12. https://doi.org/10.1002/leap.1570

Liu, R., & Shah, N. B. (2023). ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. arXiv preprint, arXiv:2306.00622 https://doi.org/10.48550/arXiv.2306.00622

Liu, J. Q., Hui, K. T., Al Zoubi, F., Zhou, Z. Z., Samartzis, D., Yu, C.C., Chang, J. R., & Wong, A. Y. (2024). The great detectives: humans versus AI detectors in catching large language model-generated medical writing. International Journal for Educational Integrity, 20(8), 1-14. https://doi.org/10.1007/s40979-024-00155-6

Meléndez, N., Gibertoni, J., Briceño, M., & Lucente, R. (2023). Metodología de evaluación cualitativa de ensayos en educación superior utilizando inteligencia artificial (IA): Modelos lingüísticos Avanzados (LLM). Actas del II Congreso de Creatividad e Innovación en Educación (CIE-2023). Presentado en II Congreso de Creatividad e Innovación en Educación. https://doi.org/10.47300/978-9962-738-17-6-11

Mondal, S., Juhi, A., Kumari, A., Dhanvijay, A., Mittal, S., & Mondal, H. (2023). Peer review in scientific publishing: Current practice, guidelines, relevancy, and way forward. Cosmoderma, 3(40). https://doi.org/10.25259/CSDM_35_2023

Mondal, H., & Mondal, S. (2024). Peer review: opportunity and challenges. Indian J Cardiovasc Dis Women, 9(2), 118-120. https://doi.org/10.25259/IJCDW_6_2024

Mostafapour, M., Fortier, J.H., Pacheco, K., Murray, H., & Garber, G.E. (2024). Evaluating Literature Reviews Conducted by Humans Versus ChatGPT: Comparative Study. JMIR AI, 3, Article e56537. https://doi.org/10.2196/56537

Nematov, D. (2025). Progress, Challenges, Threats and Prospects of ChatGPT in Science and Education: How Will AI Impact the Academic Environment? SSRN, 1-17. https://doi.org/10.2139/ssrn.5188827

Riding, J. B. (2022). An evaluation of the process of peer review. Palynology, 47(1). Article 2151052. https://doi.org/10.1080/01916122.2022.2151052

Salman, H. A., Ahmad, M. A., Ibrahim, R., & Mahmood, J. (2025). Systematic analysis of generative AI tools integration in academic research and peer review. Online Journal of Communication and Media Technologies, 15(1), Article e202502. https://doi.org/10.30935/ojcmt/15832

Shcherbiak, A., Habibnia, H., Böhm, R., & Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. Judgment and Decision Making, 19(21). https://doi.org/10.1017/jdm.2024.24

Silva, G. S., Khera, R., & Schwamm, L. H. (2024). Reviewer Experience Detecting and Judging Human Versus Artificial Intelligence Content: The Stroke Journal Essay Contest. Stroke, 55(10). 2573-2578. https://doi.org/10.1161/STROKEAHA.124.045012

Yeadon, W., Agra, E., Inyang, O., Mackay, P., & Mizouri, A. (2024). Evaluating AI and Human Authorship Quality in Academic Writing through Physics Essays. European Journal of Physics, 45, Article 055703. https://doi.org/10.1088/1361-6404/ad669d

Van Dijk, S. H. B., Brusse-Keizer, M. G. J., Bucsán, C. C., van der Palen, J., Doggen, C. J. M., & Lenferink, A. (2023). Artificial intelligence in systematic reviews: promising when appropriately used. BMJ Open, 13(7), Article e072254. https://doi.org/10.1136/bmjopen-2023-072254

Vuong, Q. H., La, V. P., Nguyen, M. H., Jin, R., Le, T. T. (2023). Are we at the start of the artificial intelligence era in academic publishing? Science Editing, 10(2), 158-164. https://doi.org/10.6087/kcse.310