

Evaluación de textos científicos universitarios asistida por IA: un estudio de tres niveles de complejidad

AI-assisted evaluation of university scientific texts: a study of three complexity levels

Elena Tzetzagary Aguirre Mejía¹, Luz María Ramírez Sandoval², Francisco Javier Ramírez Sandoval³

¹TECNM-Instituto Tecnológico de la Laguna, Blvd. Revolución, Torreón, Coahuila, México, C.P. 27000;

^{2,3}TECNM-Instituto Tecnológico de Torreón Carr. Torreón-San Pedro km 7.5 Ejido Ana. Torreón, Coahuila, México C.P. 27170

ORCID: ¹<https://orcid.org/0000-0003-4472-6025> | ²<https://orcid.org/0000-0003-3329-7798> | ³<https://orcid.org/0009-0009-2798-7048>

Autores de correspondencia: int.etaguirrem@lalaguna.tecnm.mx; posgrado.uni.05@gmail.com; fjavier.rs@gmail.com

Resumen:

La inteligencia artificial generativa (IAG) se ha convertido en una herramienta de apoyo a la escritura académica, pero su impacto en estudiantes noveles es controvertido. Objetivo: Analizar el desempeño en escritura académica de estudiantes de nivel inicial a partir de la retroalimentación generada por IA bajo una consigna estandarizada, comparando tres niveles de complejidad textual. Resultados: Se evaluaron 360 textos de 30 estudiantes. La progresión media fue ascendente (básico=74.8, metodológico=79.1, avanzado=80.2), con diferencias significativas entre el nivel básico y los otros dos ($p<.05$, $\epsilon^2=0.308$). La precisión metodológica fue la dimensión más deficiente (medias 68.3-76.9). Más del 90% presentó falta de precisión metodológica y redundancia. Limitaciones del estudio / implicaciones: Ausencia de grupo control, muestra pequeña y de una sola institución; no se midió transferencia del aprendizaje. Se requiere mediación docente para tareas de alta complejidad. Originalidad: Se emplea un prompt estructurado que evalúa siete dimensiones, revelando problemas no percibidos por los estudiantes (lenguaje coloquial, debilidad conceptual). Hallazgos / conclusiones: La IA actúa como “espejo metacognitivo”, pero es insuficiente para mejorar tareas avanzadas. Se recomienda integrar IA con retroalimentación docente y diseñar ejercicios específicos de precisión metodológica.

Palabras Clave: inteligencia artificial, escritura académica, precisión metodológica, retroalimentación automática, estudiantes principiantes.

Abstract:

Generative artificial intelligence (GAI) has become a writing support tool, but its impact on novice students is controversial. Objective: To analyze academic writing performance of first-semester students based on standardized AI-generated feedback, comparing three levels of textual complexity. Results: 360 texts from 30 students were evaluated. Mean progression was ascending (basic=74.8, methodological=79.1, advanced=80.2), with significant differences between basic and the other two ($p<.05$, $\epsilon^2=0.308$). Methodological precision was the weakest dimension (means 68.3-76.9). Over 90% showed lack of methodological precision and redundancy. Limitations / implications: No control group, small sample from one institution; learning transfer not measured. Teacher mediation is required for high-complexity tasks. Originality: A structured prompt assessing seven dimensions revealed problems not perceived by students (colloquial language, conceptual weakness). Findings / conclusions: AI acts as a “metacognitive mirror” but is insufficient for advanced tasks. We recommend integrating AI with teacher feedback and designing specific exercises for methodological precision.

Keywords: artificial intelligence, academic writing, methodological precision, automated feedback, novice students.

Introducción

La inteligencia artificial generativa (IAG) ha irrumpido en la educación superior con una velocidad sin precedentes. En apenas tres años, el uso de herramientas como ChatGPT se ha generalizado entre los estudiantes. Según el informe de Juventud del Parlamento Europeo de 2024, el 57% de los jóvenes europeos informaron haber utilizado aplicaciones basadas en IA para estudiar y escribir documentos (Raptopoulou, 2025). Esta adopción masiva ha generado un debate polarizado. Por un lado, algunos autores advierten sobre el riesgo de externalizar los procesos cognitivos superiores (Alasmari & Alshammari, 2024; Díaz-Cuevas & Rodríguez-Herrera, 2024). Por otro lado, otros defienden su potencial como andamiaje para la mejora de la redacción (Maturana, 2025).

Las evidencias empíricas disponibles son aún limitadas y, en ocasiones, contradictorias. Un estudio de Steiss et al. (2024) encontró que, si bien los estudiantes que utilizaron ChatGPT mostraron una mejora en la calidad de sus ensayos, persistían diferencias significativas en comparación con aquellos que recibieron retroalimentación humana, especialmente en tareas que requerían un razonamiento de orden superior. En contraste, una revisión sistemática de Silgado-Tuñón y López-Flores (2025) sobre el uso de IAG en educación superior concluyó que, aunque la IA facilita ciertos procesos formales de escritura, su eficacia depende críticamente de la integración pedagógica y del desarrollo de habilidades metacognitivas. Asimismo, una revisión sistemática de estudios sobre ChatGPT en escritura académica indicó que los estudiantes

utilizan estas herramientas principalmente para la corrección gramatical, la generación de ideas y la retroalimentación formativa (Lee et al., 2025).

El presente estudio se desarrolló en el contexto de una universidad pública mexicana, perteneciente al Tecnológico Nacional de México (TECNM), con estudiantes de primer semestre de la carrera de Ingeniería en Sistemas, en la asignatura de Fundamentos de Investigación. En este escenario, los alumnos carecen de formación previa en redacción científica formal, lo que convierte a este entorno en un espacio propicio para explorar el impacto de la retroalimentación asistida por IA en niveles iniciales. Si bien en América Latina la investigación sobre IAG en educación superior ha comenzado a emerger (Díaz-Cuevas & Rodríguez-Herrera, 2024; Silgado-Tuñón & López-Flores, 2025), aún existen pocos estudios que aborden de manera sistemática el análisis de errores de escritura en textos con diferente complejidad metodológica, particularmente en carreras de ingeniería.

Frente a esta paradoja, emerge un consenso en tres direcciones. Primero, se enfatiza la necesidad de desarrollar una conciencia metacrítica sobre el uso de la IAG. Alasmari y Alshammari (2024) demostraron que el uso acrítico de la IA puede llevar a una participación superficial en las tareas de escritura. Segundo, se propone la figura del escritor aumentado o la noción de "escritura centauro". Maturana (2025) postula que la IA no debe reemplazar al autor, sino transformarlo en un ente analítico que interactúa críticamente con la tecnología. Tercero, se exploran modelos de retroalimentación híbridos. Estudios recientes sobre retroalimenta-

ción con IA han mostrado que la combinación de retroalimentación generada por IA con la retroalimentación del profesor o de pares puede generar mejores resultados que el uso aislado de la IA (Banihashem et al., 2024; Escalante et al., 2023).

No obstante, a pesar de los avances documentados, persiste un vacío de conocimiento respecto a cómo la retroalimentación automatizada impacta en tareas de creciente complejidad textual, especialmente en estudiantes novatos de ingeniería en contextos latinoamericanos. Son precisamente los modelos híbridos los que ofrecen una vía pedagógica prometedora, y el presente estudio se inscribe en esa línea. Su objetivo es analizar el desempeño de la escritura académica en estudiantes de nivel inicial a partir de la retroalimentación generada por una IA bajo una consigna estandarizada, y comparar los resultados en tres niveles de complejidad textual. Las preguntas de investigación son: (1) ¿Cuáles son las puntuaciones medias en claridad, coherencia, precisión metodológica y uso de conectores en los tres niveles de escritura? (2) ¿Qué problemas de redacción detecta la IA con mayor frecuencia? (3) ¿Existe una progresión significativa en la calidad de la escritura a medida que aumenta la complejidad del texto?

Este estudio no pretende validar la IA como evaluadora, sino describir los patrones de error y las áreas de mejora que la IA hace visibles. Se busca así aportar evidencia empírica para el diseño de intervenciones docentes que integren la IAG de manera crítica y estructurada en la educación superior.

Metodología

Se realizó un estudio de enfoque mixto con diseño secuencial, de corte transversal y alcance descriptivo, con un solo grupo. Este diseño permitió, por una parte, cuantificar las puntuaciones obtenidas en las dimensiones evaluadas, y por otra, analizar cualitativamente las reflexiones metacognitivas de los estudiantes para identificar patrones de error y áreas de mejora. El contexto fue una asignatura de Fundamentos de Investigación del primer semestre de Ingeniería en Sistemas en una universidad pública mexicana, perteneciente al Tecnológico Nacional de México (TECNM). La actividad evaluada formaba parte de sesiones prácticas sobre uso ético y crítico de IA. Dado que el objetivo era describir el desempeño y los patrones de error, y no comparar condiciones experimentales, no se utilizó grupo de control, lo cual es consistente con un diseño no experimental. El estudio se ajustó a los principios éticos de la Declaración de Helsinki y fue aprobado por el comité de ética de la institución.

Participaron 30 estudiantes (18 hombres, 12 mujeres) de entre 18 y 20 años ($M = 18.7$, $DE = 0.8$). Ninguno tenía experiencia previa en redacción científica formal ni en el uso de IAG para fines académicos. Todos consintieron participar de manera voluntaria y se anonimizaron sus datos.

La variable independiente fue el nivel de complejidad de escritura, definida conceptualmente como el grado de exigencia cognitiva y metodológica requerida para la redacción del texto, y operacionalizada mediante la clasificación en tres niveles: básico (redacción científica elemental), metodoló-

gico (con hipótesis y variables explícitas) y avanzado (interpretación de resultados estadísticos). Las variables dependientes fueron las puntuaciones globales y por dimensión (claridad argumentativa, coherencia y estructura, precisión metodológica y uso de conectores), definidas operacionalmente como los puntajes asignados por la IA en una escala de 0 a 100.

Ejercicios de redacción. Cada estudiante redactó 12 textos breves (300-500 palabras). Los textos se basaron en tablas que proporcionaban título, objetivo, hipótesis (cuando correspondía), variables, metodología y resultados. Los 12 textos se organizaron originalmente en cuatro niveles de complejidad ascendente (1A, 1B, 1C; 2A, 2B, 2C; 3A, 3B, 3C; 4A, 4B, 4C). Para el análisis final, se agruparon en tres niveles: (a) básico (niveles 1A, 1B, 1C): redacción científica elemental (mapas conceptuales, aprendizaje colaborativo, microvideos); (b) metodológico (niveles 2A, 2B, 2C): redacción con hipótesis explícitas y variables (IA generativa, gamificación, estrategias metacognitivas); (c) avanzado (niveles 3A, 3B, 3C, 4A, 4B, 4C): interpretación de resultados estadísticos y redacción tipo artículo científico.

Los ejercicios de redacción fueron diseñados por un equipo de tres investigadores con experiencia en metodología de la investigación y escritura académica. La validez de contenido se estableció mediante juicio de expertos, con la participación de dos académicos externos especializados en didáctica de la escritura, quienes evaluaron la pertinencia y claridad de las tablas de datos proporcionadas a los estudiantes. El prompt de evaluación fue

sometido a un proceso de pilotaje con cinco textos de muestra, ajustando su redacción hasta lograr consistencia en las respuestas de la IA.

Prompt de evaluación. Se diseñó un prompt estandarizado que instruía a la IA (ChatGPT-4, versión 15 de enero de 2026, temperatura = 0) para que actuara como "revisor académico experto en redacción científica universitaria y metodología de investigación". El prompt solicitaba evaluar siete dimensiones en una escala de 0 a 100: claridad argumentativa, coherencia y estructura, precisión metodológica, uso de conectores, redundancia, debilidad conceptual y posible similitud textual. Además, pedía un análisis crítico de máximo 120 palabras y tres propuestas de mejora prioritaria.

Cada estudiante redactó un texto de 200 palabras respondiendo a cuatro preguntas: principal problema detectado, errores repetidos, nivel con mejores resultados y estrategias de mejora.

El procedimiento constó de cinco pasos secuenciales:

1. Redacción inicial. Los estudiantes redactaron los 12 textos en sesiones previas, con base en las tablas proporcionadas y sin ayuda de IA.
2. Evaluación con IA. Cada estudiante ingresó sus textos al chatbot junto con el prompt. La IA generó una tabla de calificaciones y un análisis crítico.
3. Sistematización. Los estudiantes completaron una tabla resumiendo las calificaciones promedio por texto y por nivel.

4. Reflexión metacognitiva. Cada estudiante redactó su reflexión de 200 palabras.

5. Recolección de datos. Se recopilaron todas las tablas y reflexiones. Se extrajeron las puntuaciones medias y se codificaron las categorías de problemas.

Variables y análisis estadístico

- Variable independiente (categórica): nivel de complejidad de escritura.
- Variables dependientes (cuantitativas): puntuación global promedio y puntuaciones por dimensión.

Se calcularon medidas de tendencia central y dispersión. Se aplicó la prueba de Shapiro-Wilk para evaluar la normalidad. Dado que los puntajes no siguieron una distribución normal ($p < .05$), se utilizó la prueba no paramétrica de Kruskal-Wallis para comparar los tres niveles. Las comparaciones post hoc se realizaron con corrección de Bonferroni. Para estimar la magnitud de las diferencias, se calculó el tamaño del efecto mediante epsilon cuadrado ($\epsilon^2 = H / (n - 1)$). Para comparar las dimensiones dentro de cada nivel se empleó la prueba de Friedman, con comparaciones post hoc de Wilcoxon ajustadas por Bonferroni. Todos los análisis cuantitativos se realizaron utilizando el software estadístico JASP (versión 0.18.3).

Se realizó un análisis de contenido de las 30 reflexiones metacognitivas, siguiendo un proceso inductivo de codificación abierta. Dos codificado-

res independientes, ambos investigadores con experiencia en análisis cualitativo y desconocedores de la identidad de los participantes, clasificaron las menciones a problemas de redacción en categorías emergentes. El proceso incluyó una primera fase de lectura flotante, seguida de codificación línea por línea y agrupación en categorías temáticas. El acuerdo intercodificador fue del 89%, medido mediante el coeficiente kappa de Cohen ($\kappa = 0.85$). No se utilizó software especializado para el análisis cualitativo, sino un proceso manual de categorización con matrices de doble entrada.

Selección de la submuestra para el análisis por dimensiones. Para el análisis detallado por dimensiones se seleccionó una submuestra de 15 estudiantes (50% del total) mediante un muestreo intencional. Los criterios de inclusión fueron: disponibilidad de textos completos en los tres niveles de complejidad y ausencia de errores de formato que pudieran afectar la evaluación por IA. Esta submuestra garantiza la representatividad de la muestra total y permite un análisis más fino de las dimensiones evaluadas.

Limitaciones metodológicas

Este estudio presenta varias limitaciones. Primera, las calificaciones provienen de la misma IA que evalúa, lo que introduce un posible sesgo de consistencia interna que se discute en el apartado correspondiente. Segunda, la muestra es pequeña y de una sola institución, lo que limita la generalización de los hallazgos. Tercera, el orden secuencial de los textos pudo confundir el efecto de la complejidad con el efecto de la práctica, aunque los

niveles se alternaron para minimizar este riesgo. Cuarta, no se midió la transferencia del aprendizaje a otras tareas de escritura. Estas limitaciones se consideran al interpretar los resultados y se sugieren líneas futuras para superarlas.

Resultados

A continuación, se presentan los resultados obtenidos a partir del análisis cuantitativo y cualitativo, siguiendo el orden de las preguntas de investigación. Se incluyen estadísticos descriptivos, pruebas de comparación entre niveles y frecuencias de los problemas de redacción detectados.

Estadísticos descriptivos por nivel

La Tabla 1 presenta las medidas de tendencia central y dispersión de las puntuaciones globales para cada nivel de complejidad. Se observa una progresión ascendente de las medias, desde el nivel básico ($M = 74.83$) hasta el avanzado ($M = 80.23$). La desviación estándar es mayor en el nivel básico ($DE = 11.21$), lo que indica una alta heterogeneidad inicial en el desempeño de los estudiantes. En los niveles metodológico ($DE = 9.42$) y avanzado ($DE = 9.82$), la dispersión disminuye, lo que sugiere una mayor homogeneidad en el rendimiento a medida que aumenta la complejidad de la tarea.

Tabla 1. Estadísticas descriptivas de las puntuaciones globales por nivel de complejidad ($N = 30$)

Nivel	Media	Mediana	DE	Mínimo	Máximo
Básico	74.83	76.00	11.21	43.0	89.7
Metodológico	79.13	81.00	9.42	50.0	91.3
Avanzado	80.23	83.50	9.82	56.0	93.6

Fuente: Elaboración propia.

Para complementar el análisis descriptivo y visualizar el comportamiento de la variable nivel de complejidad en escritura, la Figura 1 presenta la distribución de las puntuaciones globales mediante un diagrama de cajas (boxplot) para cada nivel. En este gráfico se aprecia claramente que la mediana del nivel básico (76.00) es inferior a la de los otros dos niveles, y que el rango intercuartílico es más amplio en el nivel básico, confirmando la mayor variabilidad inicial.

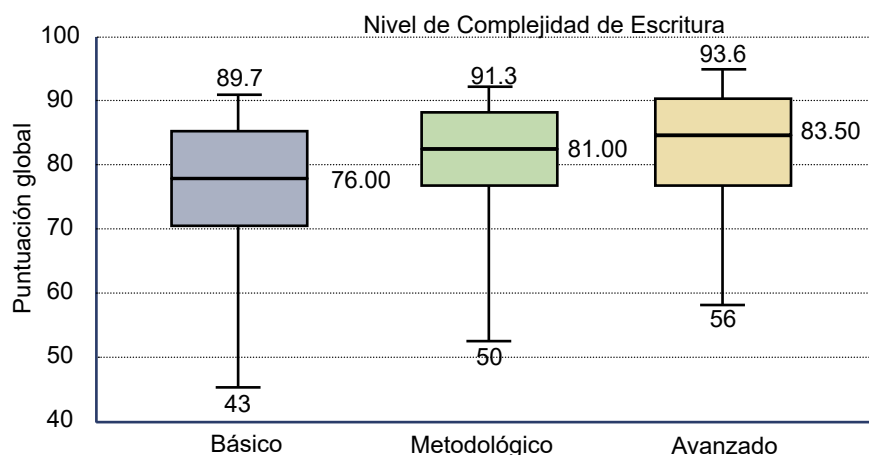


Figura 1. Distribución de frecuencias de las puntuaciones globales por nivel de complejidad.

Nota: La línea central representa la mediana, la caja muestra el rango intercuartílico ($Q1-Q3$) y los bigotes indican los valores mínimo y máximo.

Fuente: Elaboración propia a partir de datos de estudio.

Para determinar si las diferencias observadas en las medias son estadísticamente significativas, se aplicó la prueba no paramétrica de Kruskal-Wallis, dada la ausencia de normalidad en los puntajes (Shapiro-Wilk, $p < .05$). Los resultados mostraron diferencias significativas entre los tres niveles de complejidad, $H(2) = 8.94$, $p = .011$. El tamaño del efecto fue grande, $\epsilon^2 = 0.308$, lo que respalda la relevancia práctica de esta diferencia en el contexto educativo.

Las comparaciones post hoc con corrección de Bonferroni revelaron que el nivel básico ($Mdn = 76.00$) fue significativamente inferior al nivel metodológico ($Mdn = 81.00$), $p = .008$, y al nivel avanzado ($Mdn = 83.50$), $p = .004$. En contraste, no se encontró diferencia significativa entre el nivel metodológico y el avanzado, $p = .312$. Estos hallazgos

indican que el salto cualitativo más importante en la calidad de la escritura ocurre al incorporar hipótesis y variables explícitas, mientras que la transición hacia tareas de interpretación estadística no representa un avance significativo adicional en la puntuación global.

Desempeño por dimensiones

Para el análisis detallado de las dimensiones evaluadas (claridad argumentativa, coherencia y estructura, precisión metodológica y uso de conectores), se seleccionó una submuestra de 15 estudiantes (50% del total), cuyos criterios de inclusión se describen en el apartado de metodología. La Tabla 2 muestra las puntuaciones medias obtenidas en cada dimensión para los tres niveles.

Tabla 2. Puntuaciones medias por dimensión y nivel de complejidad ($n = 15$)

Dimensión	Básico	Metodológico	Avanzado
Claridad argumentativa	78.2	82.5	84.1
Coherencia y estructura	76.5	81.0	83.7
Precisión metodológica	68.3	74.8	76.9
Uso de conectores	72.1	76.5	79.2

Fuente: Elaboración propia.

De manera consistente en los tres niveles, la precisión metodológica resultó ser la dimensión peor valorada, con medias que oscilan entre 68.3 (básico) y 76.9 (avanzado). Para evaluar la significación de estas diferencias entre dimensiones dentro de cada nivel, se aplicó la prueba de Friedman. En los tres niveles se encontraron diferencias significativas (básico: $p < .001$; metodológico: $p < .001$; avanzado: $p < .001$). Las comparaciones post hoc

con corrección de Bonferroni indicaron que la precisión metodológica fue significativamente inferior a las otras tres dimensiones en todos los niveles ($p < .01$ en todos los casos). Este patrón sugiere que la dificultad para describir instrumentos, procedimientos y análisis estadísticos constituye una debilidad estructural en la escritura de los estudiantes principiantes, independientemente de la complejidad del texto.

Problemas de redacción detectados

La Tabla 3 presenta la frecuencia de ocho categorías de problemas, identificadas a partir de las reflexiones metacognitivas de los estudiantes y los análisis críticos generados por la IA. Más del 90% de los estudiantes presentó falta de precisión metodológica (93%) y redundancia o repetición de ideas

(87%). El uso limitado o inadecuado de conectores afectó al 80% de los participantes, mientras que el lenguaje coloquial o informal persistió en tres cuartas partes de los textos (73%). La debilidad conceptual, evidenciada en la ausencia de definiciones o referencias teóricas, se presentó en el 67% de los casos.

Tabla 3. Frecuencia de problemas de redacción detectados por la IA (N = 30)

Problema	n	%
Falta de precisión metodológica	28	93
Redundancia / repetición de ideas	26	87
Uso limitado o inadecuado de conectores	24	80
Lenguaje coloquial o informal	22	73
Debilidad conceptual (sin definiciones ni referencias)	20	67
Omisión de detalles metodológicos	18	60
Riesgo de similitud textual por frases genéricas	15	50
Errores gramaticales / puntuación	14	47

Fuente: Elaboración propia.

Las siguientes citas textuales de los estudiantes ilustran de manera representativa estos problemas:

E05: "El principal problema fue la falta de precisión metodológica. En varios ejercicios, la descripción de los instrumentos de evaluación era insuficiente."

E08: "Se repitieron bastantes problemas de cohesión entre ideas, uso limitado de conectores y repetición de ideas."

E16: "Expresiones como 'o sea' o afirmaciones en primera persona del plural interrumpieron la coherencia del discurso académico."

E12: "El término 'autonomía académica' no se fundamenta con autores. Lo usé sin saber bien qué significaba."

Estos testimonios refuerzan los hallazgos cuantitativos y evidencian que la retroalimentación generada por IA permite a los estudiantes tomar conciencia de aspectos de su escritura que previamente pasaban desapercibidos, sentando las bases para la discusión de los hallazgos y sus implicaciones pedagógicas.

Conclusiones

Este estudio evaluó la calidad de 360 textos científicos producidos por 30 estudiantes de nivel inicial,

utilizando un prompt estandarizado de IA. A partir de los resultados obtenidos, se derivan cinco hallazgos principales.

Primero, se confirma una progresión significativa en la calidad de escritura únicamente entre el nivel básico ($M = 74.8$, $Mdn = 76.0$) y el metodológico ($M = 79.1$, $Mdn = 81.0$), con diferencias estadísticamente significativas ($*p* = .008$). En contraste, la diferencia entre el nivel metodológico y el avanzado ($M = 80.2$, $Mdn = 83.5$) no resultó significativa ($*p* = .312$), lo que indica que el salto cualitativo más relevante ocurre al incorporar hipótesis y variables explícitas. El tamaño del efecto obtenido ($\epsilon^2 = 0.308$) respalda la relevancia práctica de esta diferencia en el contexto de la formación inicial en investigación.

Segundo, la precisión metodológica se confirma como la dimensión más deficiente en todos los niveles evaluados, con medias que oscilan entre 68.3 y 76.9. Más del 90% de los estudiantes presentó problemas para describir instrumentos, procedimientos y análisis estadísticos. Este hallazgo supera en magnitud lo reportado en estudiantes de semestres superiores (Montolío, 2014) y refleja el nivel inicial de los participantes, lo que subraya la urgencia de intervenciones tempranas en esta área.

Tercero, la redundancia (87%) y el uso limitado o inadecuado de conectores (80%) constituyen problemas transversales que afectan la cohesión textual. El lenguaje coloquial o informal persiste en tres cuartas partes de los estudiantes, lo que coincide con estudios previos sobre escritura de principiantes (Wilson & Czik, 2024) y sugiere la necesidad de talleres específicos de alfabetización académica desde el primer semestre.

Cuarto, la evaluación asistida por IA, mediante un prompt estructurado y con parámetros fijos, demostró ser útil para detectar problemas que los estudiantes no perciben por sí mismos. Los participantes valoraron la retroalimentación como un "espejo" que les mostró sus lagunas metodológicas y su estilo coloquial. Este resultado concuerda con investigaciones que muestran que la retroalimentación generada por IA puede aumentar la conciencia metacognitiva (Banihashem et al., 2024; Escalante et al., 2023), siempre que se acompañe de una reflexión guiada.

Quinto, la ausencia de mejora significativa en el nivel avanzado sugiere que la IA, utilizada de manera aislada, es insuficiente para promover saltos cualitativos en tareas de alta complejidad interpretativa. Este hallazgo respalda la necesidad de mediación docente y de modelos híbridos de retroalimentación, en línea con estudios que abogan por la combinación de IA con la revisión por pares o la tutoría humana (Chan et al., 2024; Steiss et al., 2024).

Discusiones

Los hallazgos de este estudio confirman que la evaluación asistida por IA, mediante un prompt estructurado, puede funcionar como una herramienta de retroalimentación formativa que revela problemas de escritura no percibidos por los estudiantes principiantes. La progresión significativa observada entre el nivel básico y el metodológico ($p = .008$) sugiere que la IA es particularmente efectiva para detectar deficiencias en textos que requieren una estructura metodológica explícita, como la inclusión

de hipótesis y variables. No obstante, la ausencia de mejora significativa entre el nivel metodológico y el avanzado ($p = .312$) indica que la retroalimentación automatizada, utilizada de manera aislada, es insuficiente para promover saltos cualitativos en tareas de alta complejidad interpretativa. Este resultado coincide con lo reportado por Steiss et al. (2024), quienes encontraron que, si bien los estudiantes que utilizaron ChatGPT mostraron mejoras en la calidad de sus ensayos, persistían diferencias significativas en comparación con aquellos que recibieron retroalimentación humana, especialmente en tareas que requerían razonamiento de orden superior.

La deficiencia sistemática en precisión metodológica, con medias que oscilan entre 68.3 y 76.9 en los tres niveles, constituye el hallazgo más relevante del estudio. Más del 90% de los estudiantes presentó problemas para describir instrumentos, procedimientos y análisis estadísticos, lo que supera en magnitud lo reportado en estudiantes de semestres superiores (Montolío, 2014) y refleja el nivel inicial de los participantes. Esta debilidad estructural, independiente de la complejidad del texto, sugiere que la escritura metodológica requiere un desarrollo cognitivo y práctico que no se resuelve únicamente con retroalimentación automatizada. En este sentido, los resultados respaldan la postura de Silgado-Tuñón y López-Flores (2025), quienes concluyeron que la eficacia de la IA depende críticamente de la integración pedagógica y del desarrollo de habilidades metacognitivas.

Un aspecto que merece atención es la alta frecuencia de redundancia (87%) y uso limitado de conec-

tores (80%), así como la persistencia del lenguaje coloquial en tres cuartas partes de los estudiantes. Estos patrones coinciden con estudios previos sobre escritura de principiantes (Wilson & Czik, 2024) y sugieren que los estudiantes transfieren a la escritura académica las convenciones del lenguaje oral y coloquial, sin conciencia plena de las exigencias del registro formal. La retroalimentación generada por IA logró hacer visibles estos problemas, como lo evidencian las reflexiones metacognitivas de los participantes, quienes valoraron la evaluación como un "espejo" que les mostró lagunas que previamente desconocían. Este hallazgo concuerda con investigaciones que muestran que la retroalimentación generada por IA puede aumentar la conciencia metacognitiva (Banihashem et al., 2024; Escalante et al., 2023), siempre que se acompañe de una reflexión guiada y de un andamiaje docente que permita a los estudiantes procesar y aplicar las sugerencias recibidas.

Sin embargo, es necesario reconocer que la dependencia de una única fuente de evaluación —la misma IA que genera las puntuaciones— constituye una limitación metodológica relevante. Si bien el prompt estandarizado y los parámetros fijos (temperatura = 0) buscaban garantizar la consistencia de las respuestas, no se realizó una triangulación con evaluadores humanos que permitiera validar externamente las calificaciones asignadas por la IA. Esta decisión metodológica, aunque coherente con el objetivo descriptivo del estudio, introduce un posible sesgo de consistencia interna que debe ser considerado al interpretar los resultados. Futuras investigaciones deberían incorporar la comparación con juicios de expertos humanos para determinar el

grado de concordancia y la validez de la evaluación automatizada en diferentes dimensiones de la escritura académica.

La ausencia de mejora en el nivel avanzado sugiere que la IA, en su estado actual, es insuficiente para abordar tareas que requieren interpretación de resultados estadísticos y redacción tipo artículo científico. Este hallazgo respalda la necesidad de modelos híbridos de retroalimentación, en línea con lo propuesto por Chan et al. (2024) y Steiss et al. (2024), que combinan la retroalimentación generada por IA con la revisión por pares o la tutoría humana. En este sentido, la IA podría funcionar como una primera capa de retroalimentación que identifica problemas formales y estructurales, mientras que el docente o los pares podrían enfocarse en aspectos de orden superior como la argumentación, la profundidad conceptual y la coherencia global del texto.

En el contexto latinoamericano, donde la investigación sobre IAG en educación superior es aún incipiente (Díaz-Cuevas & Rodríguez-Herrera, 2024; Silgado-Tuñón & López-Flores, 2025), este estudio aporta evidencia empírica sobre los patrones de error y las áreas de mejora que la IA hace visibles en estudiantes de ingeniería de primer ingreso. Los hallazgos subrayan la necesidad de intervenciones tempranas y específicas en precisión metodológica, así como la importancia de diseñar ejercicios que aborden de manera explícita las dimensiones más deficientes detectadas. La integración de la IA en la enseñanza de la escritura académica no debe ser vista como un sustituto de la mediación docente, sino como un andamiaje que, utilizado crítica-

mente, puede potenciar el desarrollo de habilidades metacognitivas y la conciencia sobre los propios procesos de escritura.

Por lo tanto, los resultados de este estudio sugieren que la evaluación asistida por IA puede ser una herramienta valiosa para la retroalimentación formativa en estudiantes principiantes, siempre que se utilice de manera complementaria a la intervención docente y se acompañe de procesos de reflexión guiada. La precisión metodológica emerge como la dimensión más crítica y demanda una atención pedagógica prioritaria desde los primeros semestres. Las futuras investigaciones deberían explorar modelos híbridos de retroalimentación y evaluar la sostenibilidad de las mejoras en el tiempo mediante diseños longitudinales.

Recomendaciones para la práctica docente

- Diseñar ejercicios específicos de precisión metodológica que incluyan la obligación de nombrar instrumentos, reportar validez y confiabilidad, especificar duración de intervenciones y detallar pruebas estadísticas.
- Implementar talleres de eliminación de redundancias usando ejemplos anonimizados de los propios estudiantes para que reescriban textos en la mitad de palabras.
- Crear ejercicios de conectores académicos con listas categorizadas (adición, contraste, causa-efecto, conclusión) y exigir variedad en su uso.
- Promover el uso combinado de IA y revisión por

pares: después de la retroalimentación de la IA, los estudiantes discuten en parejas cómo incorporar las sugerencias.

- Enseñar a referenciar incluyendo al menos una cita obligatoria por párrafo en textos de nivel avanzado, con verificación de normas APA.

Futuras líneas de investigación

Se necesitan estudios longitudinales con mediciones repetidas para evaluar si la práctica continua con retroalimentación de IA produce mejoras sostenidas en el tiempo. También sería valioso comparar grupos con y sin mediación docente para aislar el efecto de la IA sola y determinar el valor agregado de la intervención humana. Además, se podría explorar la calidad de la retroalimentación generada por diferentes modelos (ChatGPT-4, Gemini, Copilot) y su aceptación por parte de los estudiantes, así como incorporar medidas de esfuerzo cognitivo para contrastar hipótesis de externalización de procesos cognitivos. Finalmente, se sugiere ampliar la muestra a otras instituciones y carreras para evaluar la generalización de los hallazgos.

Referencias

- Alasmari, T. M., & Alshammari, A. S. (2024). Assessing the role of ChatGPT in improving Saudi EFL students' academic writing performance. *Journal of Educational Computing Research*, 62(5), 1289-1312.
- Banihashem, S. K., Taghizadeh Kerman, N., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: Peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, 21, Article 23. <https://doi.org/10.1186/s41239-024-00455-4>
- Cassany, D. (2006). Taller de textos: Leer, escribir y comentar en el aula. *Paidós*.
- Chan, S., Lo, N., & Wong, A. (2024). Generative AI and essay writing: Impacts of automated feedback on revision performance and engagement. *rEFLections*, 31(3), 1249-1284.
- Díaz-Cuevas, A. P., & Rodríguez-Herrera, J. D. (2024). Usos de la Inteligencia Artificial en la escritura académica: experiencias de estudiantes universitarios en 2023. *Cuaderno de Pedagogía Universitaria*, 21(42), 25-44. <https://doi.org/10.29197/cpu.v21i42.595>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20, Article 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Frome, J. (2024, June 15-18). Content analysis of student AI use in a first-year writing course [Paper presentation]. *Higher Education Conference in Singapore (HECS) 2024*, National University of Singapore.
- Lee, H.-Y., Huang, Y.-M., & Wu, T.-T. (2025). ChatGPT in education: A systematic review of current landscape, limitations and future directions through general system theory lens. *European Journal of Education*, 60(4). <https://doi.org/10.1111/ejed.70262>

- Le, T. T. P., Quach, T. T. N., Pham, T. X. T., & Nguyen, D. L. (2025). A systematic review of how ChatGPT is perceived and utilized in EFL writing classes: Implications for Vietnam. *AsiaCALL Online Journal*, 16(1), 295-311.
- Maturana, A. J. (2025). Inteligencias Artificiales Generativas y prácticas de escritura académica en la Educación Superior: Un estado del arte desde aportes publicados en América Latina en 2022-2023. *Revista RAES*, 17(30), 98-113.
- Montolío, E. (2014). Manual práctico de escritura académica (3ª ed.). *Ariel*.
- Raptopoulou, A. (2025). Generative AI in European higher education: Youth perspectives and usage patterns. *European Journal of Education*, 60(3). <https://doi.org/10.1111/ejed.70131>
- Silgado-Tuñón, D. A., & López-Flores, J. I. (2025). Inteligencia Artificial Generativa en la Educación Superior: una Revisión Sistemática. UNION: *Revista Iberoamericana de Educación Matemática*, 36(1), 45-68.
- Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. *Learning and Instruction*, 91, 101894.
- Wilson, J., & Czik, A. (2024). Automated writing evaluation and feedback: A systematic review of effectiveness. *Computers and Education*, 193, 104671.
- Yang, H., Gao, C., & Shen, H. (2024). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies*, 29, 3837-3858.
- Yan, D. (2023). Impact of ChatGPT on EFL students' academic writing skills: A mixed-methods intervention study. *Smart Learning Environments*, 10, Article 8. <https://doi.org/10.1186/s40561-023-00295-5>
- Zhai, X. (2024). Transforming education with AI: A systematic review of ChatGPT's role in learning, academic practices, and institutional adoption. *Heliyon*, 10(4), e26055. <https://doi.org/10.1016/j.heliyon.2024.e26055>
- Zhang, L. (2024). Human-AI collaboration patterns in AI-assisted academic writing. *Studies in Higher Education*, 49(5), 889-905.