JOURNAL OF EDUCATION, RESEARCH AND INNOVATION

TIPO: Artículo original
PUBLICADO: 13 de Octubre, 2026
Volumen: 1
Número: 1
DOI: https://doi.org/10.5281/zenodo.23354388

Retroalimentación apoyada por inteligencia artificial y escritura argumentativa en estudiantes universitarios

Artificial intelligence supported feedback and argumentative writing among university students

Feedback apoiado por inteligência artificial e escrita argumentativa em estudantes universitários

Damián Pérez Guillerme[1]
Lila Maribel Morán Borja[2]
Johana del Carmen Parreño Sánchez[3]
Tanya Del Rocío Sánchez Salazar[4]

[1]Licenciado en Educación graduado de la Universidad de Ciencias Pedagógicas Enrique José Varona, (Cuba). Especialidad Licenciatura en Educación Primaria graduado de la Universidad de Ciencias pedagógicas Enrique José Varona, (Cuba). Máster en Didáctica de las Humanidades graduado de la Universidad de Ciencias Pedagógicas Enrique José Varona, (Cuba). Doctor en Ciencias Pedagógicas graduado de la Universidad de Ciencias Pedagógicas Enrique José Varona, (Cuba).
[2]Doctora en Educación, graduada de la Universidad César Vallejo, (Perú). Magíster en Desarrollo Educativo, graduada de la Universidad Técnica de Babahoyo, (Ecuador). Licenciada en Ciencias de la Educación, mención Educación Primaria, graduada de la Universidad Técnica de Babahoyo, (Ecuador).
[3]Doctora en Educación, graduada de la Universidad César Vallejo, (Perú). Magíster en Educación Informática, graduada de la Universidad de Guayaquil, (Ecuador). Magíster en Gerencia de Proyectos Educativos y Sociales, graduada de la Universidad Técnica de Babahoyo, (Ecuador). Especialista en Liderazgo y Gerencia, graduada de la Universidad Técnica de Babahoyo, (Ecuador).
[4]Magíster en Docencia y Gerencia en Educación Superior graduada de la Universidad de Guayaquil, (Ecuador). Magíster en Educación Básica graduada de la Universidad Bolivariana del Ecuador, (Ecuador). Doctorado en Pensamiento Complejo y Construcción Transdisciplinaria del Conocimiento graduada de la Universidad Multiversidad Mundo Real Edgar Morin, (México). Posdoctorado en Investigación y Metodología Científica: Mención Especialista en Investigación y Metodología Científica graduada de la Universidad de Oriente, (México). Doctor of Philosophy (PhD) Major in complex Chinking & Transdisciplinary Construction of Knowledge, Universidad de California, (Estados Unidos).

Resumen

La calidad de un texto revisado con ayuda no demuestra por sí sola aprendizaje transferible. El objetivo fue comparar una modalidad de retroalimentación apoyada por inteligencia artificial y revisada por docentes con retroalimentación docente, utilizando pruebas de escritura argumentativa sin asistencia. Se analizó una base de 480 registros, distribuida en 24 paralelos y 12 pares, con dos mediciones basales, una evaluación posterior a seis semanas y un seguimiento cuatro semanas después. La puntuación se obtuvo mediante cinco criterios de una rúbrica de 0 a 20 y dos evaluaciones por texto. Se aplicó ANCOVA con ajuste por promedio basal, forma de evaluación y par, errores robustos CR2 por paralelo y 100 imputaciones múltiples. Se dispuso de 442 evaluaciones inmediatas y 408 de seguimiento. La diferencia ajustada inmediata fue de 1,61 puntos a favor de la modalidad asistida (IC del 95 %: 0,96 a 2,25; p < 0,001), equivalente a 0,70 desviaciones estándar basales. A cuatro semanas fue de 0,95 puntos (IC del 95 %: 0,58 a 1,32; p < 0,001). Al reducir en dos puntos las notas imputadas de la intervención y aumentar en dos las de comparación, el intervalo del seguimiento incluyó el cero. La mediana del tiempo docente por ciclo fue de 4,96 frente a 7,23 minutos. Los resultados ilustraron una ventaja ajustada inmediata y una retención más sensible a los supuestos de ausencia. La interpretación quedó limitada por la procedencia de los datos y por la estructura no aleatorizada del diseño.

Abstract

The quality of a revised text produced with assistance does not, by itself, demonstrate transferable learning. The objective was to compare artificial intelligence supported, teacher-reviewed feedback with teacher feedback using unassisted argumentative writing assessments. A dataset of 480 records, distributed across 24 classes and 12 matched pairs, was analyzed with two baseline assessments, a post-assessment after six weeks, and a follow-up four weeks later. Scores were based on five rubric criteria, totaling 0–20 points, with two ratings per text. ANCOVA adjusted for mean baseline score, assessment form, and matched pair; CR2 class-clustered standard errors and 100 multiple imputations were used. There were 442 available post-assessments and 408 follow-up assessments. The adjusted immediate difference favored the supported condition by 1.61 points (95% CI: 0.96–2.25; p < 0.001), equivalent to 0.70 baseline standard deviations. At four weeks, the difference was 0.95 points (95% CI: 0.58–1.32; p < 0.001). When imputed intervention scores were reduced by two points and comparison scores increased by two, the follow-up confidence interval included zero. Median teacher time per feedback cycle was 4.96 versus 7.23 minutes. The findings illustrated an adjusted immediate advantage and retention that was more sensitive to missing-data assumptions. Interpretation remained limited by data provenance and the nonrandomized design structure.

Resumo

A qualidade de um texto revisado com auxílio não demonstra, por si só, aprendizagem transferível. O objetivo foi comparar uma modalidade de feedback apoiado por inteligência artificial e revisado por docentes com feedback docente, utilizando avaliações de escrita argumentativa sem assistência. Foi analisada uma base de 480 registros, distribuídos em 24 turmas e 12 pares, com duas avaliações iniciais, uma avaliação após seis semanas e acompanhamento quatro semanas depois. A pontuação foi obtida por cinco critérios de uma rubrica de 0 a 20 e duas avaliações por texto. Foi aplicada ANCOVA com ajuste pela média inicial, forma de avaliação e par, erros robustos CR2 por turma e 100 imputações múltiplas. Estavam disponíveis 442 avaliações imediatas e 408 de acompanhamento. A diferença ajustada imediata favoreceu a modalidade assistida em 1,61 ponto (IC de 95%: 0,96–2,25; p < 0,001), equivalente a 0,70 desvio-padrão inicial. Após quatro semanas, a diferença foi de 0,95 ponto (IC de 95%: 0,58–1,32; p < 0,001). Quando as notas imputadas da intervenção foram reduzidas em dois pontos e as da comparação aumentadas em dois, o intervalo do acompanhamento incluiu zero. A mediana do tempo docente por ciclo foi de 4,96 frente a 7,23 minutos. Os resultados ilustraram uma vantagem ajustada imediata e retenção mais sensível aos pressupostos sobre dados ausentes. A interpretação permaneceu limitada pela origem dos dados e pela estrutura não aleatória do desenho.

PALABRAS CLAVES: *Retroalimentación formativa; Inteligencia artificial generativa; Escritura argumentativa; Transferencia del aprendizaje.*
KEYWORDS: *Formative feedback; Generative artificial intelligence; Argumentative writing; Transfer of learning.*
PALAVRAS-CHAVE: *Retroalimentação formativa; Inteligência artificial generativa; Escritura argumentativa; Transferência de aprendizagem.*

OPEN ACCESS

CORRESPONDENCIA
Damián Pérez Guillerme
damianperez82@gmail.com

[1]Universidad de Ciencias Pedagógicas Enrique José Varona, (Cuba).
[1]damianperez82@gmail.com
[2]Universidad Técnica de Babahoyo, (Ecuador).
[2]lmmoran@utb.edu.ec
[3]Universidad Técnica de Babahoyo, (Ecuador).
[3]jparreno@utb.edu.ec
[4]Universidad Técnica de Babahoyo, (Ecuador).
[4]trsanchezs@utb.edu.ec

RECIBIDO: 10 de Agosto, 2026
REVISADO: 20 de Agosto, 2026
APROBADO: 8 de Septiembre, 2026
PUBLICADO: 13 de Octubre, 2026

CITACIÓN SUGERIDA

Pérez, D., Morán, L., Parreño, J. & Sánchez, T. (2026). Retroalimentación apoyada por inteligencia artificial y escritura argumentativa en estudiantes universitarios. *CENIT Internacional Journal of Education, Research and Innovation.* 1(1). 41 - 53 https://doi.org/10.5281/zenodo.23354388

COPYRIGHT

Este artículo se publica en acceso abierto bajo los términos de la licencia CC BY-NC 4.0 Atribución/ Reconocimiento-NoComercial 4.0 Internacional. Los contenidos publicados en esta revista pueden copiarse, compartirse y adaptarse con fines no comerciales, siempre que se reconozcan debidamente la autoría y la fuente original, se proporcione un enlace a la licencia y se indiquen las modificaciones realizadas. Su utilización no debe sugerir el respaldo de los autores o de la revista a las adaptaciones o usos derivados. Estas libertades son irrevocables mientras se cumplan los términos de la licencia. Asimismo, no podrán imponerse restricciones legales ni medidas tecnológicas que limiten los derechos de uso concedidos.

Introducción

La escritura argumentativa ocupa un lugar central en la formación universitaria porque exige formular una posición, seleccionar evidencia pertinente y responder a interpretaciones alternativas. Estas operaciones requieren algo más que producir un texto gramaticalmente correcto: implican justificar las relaciones entre afirmaciones, fuentes y conclusiones. En la enseñanza de la escritura, la revisión constituye una oportunidad para hacer visibles esas decisiones. El metaanálisis de Graham y Perin (2007), centrado en adolescentes, identificó aportes de la instrucción estratégica y de diversas prácticas de escritura. Aunque sus estimaciones no se trasladan directamente a la universidad, fundamenta el interés por intervenciones que enseñen a planificar, revisar y justificar cambios.

La retroalimentación puede apoyar ese proceso cuando informa sobre la distancia entre el desempeño actual y los criterios de calidad. Hattie y Timperley (2007) distinguieron niveles de información relativos a la tarea, los procesos y la autorregulación, mostrando que no todos los comentarios cumplen la misma función. Shute (2008) destacó la relevancia de una información específica y utilizable, cuya eficacia depende también de la tarea y del receptor. En escritura, señalar que un argumento es débil resulta insuficiente si el estudiante no comprende qué evidencia falta o por qué una inferencia excede lo que permiten las fuentes.

El potencial de la retroalimentación tampoco equivale a un beneficio uniforme. Wisniewski et al. (2020) documentaron heterogeneidad entre intervenciones educativas y señalaron la importancia del contenido informativo de los mensajes. Una devolución puede orientar la revisión, saturar al estudiante con observaciones dispersas o concentrar la atención en aspectos superficiales. Por ello, comparar modalidades de retroalimentación requiere mantener objetivos y criterios comunes y describir qué cambia realmente entre condiciones. El volumen de comentarios, la velocidad de entrega o la aceptación subjetiva no sustituyen una medida del aprendizaje alcanzado.

La perspectiva autorregulatoria desplaza parte de la atención hacia lo que el estudiante hace con la información recibida. Nicol y Macfarlane-Dick (2006) vincularon la evaluación formativa con la capacidad de juzgar el propio trabajo y regular las acciones posteriores. Carless y Boud (2018) desarrollaron el concepto de alfabetización en retroalimentación para explicar las capacidades necesarias para comprender, valorar y utilizar los comentarios. Desde este enfoque, una intervención no termina cuando el docente entrega una observación: incluye la interpretación del mensaje, la decisión de aceptarlo o cuestionarlo y la revisión fundamentada del texto.

La participación activa del receptor es especialmente relevante cuando dispone de muchas sugerencias. Winstone et al. (2017) sistematizaron procesos que favorecen la implicación del estudiante con la retroalimentación y las barreras que pueden impedir su uso. Panadero et al. (2017) encontraron, mediante metaanálisis, beneficios de la autoevaluación para la autorregulación y la autoeficacia. Estas contribuciones respaldan incorporar registros de revisión y justificaciones breves de los cambios. Sin embargo, tales registros son indicadores de proceso y no demuestran, por sí solos, que el estudiante pueda resolver una nueva tarea sin apoyo.

La investigación sobre retroalimentación entre pares aporta un antecedente para separar el origen del comentario de su función pedagógica. Huisman et al. (2019) sintetizaron sus efectos sobre la escritura académica en educación superior y mostraron su potencial formativo. Este resultado no implica que cualquier agente pueda sustituir al docente en cualquier contexto. Sugiere, más bien, que la eficacia depende de los criterios utilizados, las oportunidades de revisión y la forma en que se organiza la participación. Una comparación activa debe asegurar que ambos grupos reciban una intervención docente plausible y oportunidades equivalentes de aprendizaje.

La retroalimentación automatizada precedió a los actuales modelos generativos. Stevenson y Phakiti (2014) distinguieron entre mejoras en los textos elaborados con evaluación automática y evidencia de avances más generales en competencia escrita. Fleckenstein et al. (2023) encontraron un efecto favorable promedio de la retroalimentación automatizada en un metaanálisis multinivel, con variaciones entre condiciones. Estas investigaciones justifican estudiar la automatización, pero también delimitan su alcance: sistemas, tareas y comparadores diferentes no representan una misma intervención, y una revisión mejorada no equivale necesariamente a transferencia.

Los modelos de lenguaje ampliaron las posibilidades de producir preguntas, explicaciones y comentarios adaptados a un borrador. Kasneci et al. (2023) analizaron oportunidades educativas de estas herramientas junto con problemas de exactitud, dependencia y uso crítico. Lo (2023), en una revisión rápida de la literatura inicial sobre ChatGPT, describió aplicaciones potenciales y preocupaciones sobre la fiabilidad de las respuestas. Estos trabajos aportan marcos para el diseño, pero no constituyen evidencia experimental de que una modalidad concreta mejore la escritura universitaria en español. La fluidez de una respuesta debe distinguirse de su pertinencia pedagógica y de su respaldo documental.

La comparación directa de comentarios permite examinar una parte de esa cadena. Steiss et al. (2024) evaluaron retroalimentación humana y de ChatGPT sobre textos de estudiantes de secundaria y encontraron ventajas de evaluadores humanos entrenados en varias dimensiones de calidad. El estudio no evaluó retención universitaria a largo plazo; su contribución fue mostrar que la calidad del comentario requiere criterios explícitos. Para un uso docente, esto respalda revisar la exactitud, la priorización de problemas y la orientación ofrecida antes de entregar observaciones al estudiante.

Meyer et al. (2024) evaluaron retroalimentación generada por un modelo de lenguaje en estudiantes de secundaria que escribían en inglés como lengua extranjera. La comparación con revisión sin retroalimentación mostró mejoras en la revisión del texto y en indicadores motivacionales. Su diseño responde a una pregunta distinta de comparar una modalidad asistida con comentarios elaborados por un docente y evaluar después una tarea nueva. La diferencia entre comparación pasiva y activa resulta decisiva: el beneficio de recibir comentarios no identifica automáticamente un beneficio adicional atribuible al componente generativo.

La evidencia procedente de otros dominios ayuda a delimitar hipótesis sin reemplazar la evaluación contextual. Kestin et al. (2025) encontraron resultados favorables de una tutoría con IA diseñada pedagógicamente en un curso universitario de física. Bastani et al. (2025), en matemáticas de secundaria, mostraron que el diseño del apoyo importaba para distinguir el rendimiento durante la práctica del desempeño posterior sin acceso a la herramienta. Estos hallazgos no establecen efectos esperados para la escritura en Ecuador. Sí justifican restringir la sustitución de respuestas, incorporar preguntas orientadoras y medir el desempeño autónomo.

La dimensión operativa también requiere evidencia independiente. Noy y Zhang (2023) documentaron mejoras de productividad en tareas profesionales de escritura con asistencia generativa. Ese resultado no equivale a aprendizaje estudiantil ni garantiza ahorro docente cuando los comentarios deben verificarse. La generación puede reducir el tiempo de elaboración inicial y, a la vez, introducir trabajo de comprobación o corrección. Por ello, una evaluación educativa debe registrar el tiempo total del ciclo y la presencia de errores, en lugar de atribuir eficiencia a la velocidad de respuesta del sistema.

La integración institucional supone además decisiones sobre responsabilidad y trazabilidad. Memarian y Doleck (2023) revisaron los problemas de equidad, rendición de cuentas, transparencia y ética de la IA en educación superior. Yan et al. (2024) identificaron dificultades prácticas y éticas de los modelos de lenguaje, incluidas limitaciones de reproducibilidad. En una modalidad de retroalimentación, estas consideraciones se concretan en conservar instrucciones y salidas, proteger los textos y permitir que el docente corrija o rechace sugerencias. El uso de una rúbrica común no elimina la necesidad de supervisión ni convierte automáticamente la intervención en equitativa.

La pregunta que articuló este estudio fue si una modalidad de retroalimentación generativa revisada por docentes se asociaba con una mayor puntuación de escritura argumentativa autónoma que una modalidad de retroalimentación docente con tareas, criterios y tiempos de trabajo estudiantil equivalentes. El objetivo principal fue estimar la diferencia ajustada en una prueba nueva inmediatamente posterior a seis semanas de trabajo. Los objetivos secundarios fueron examinar el seguimiento a cuatro semanas, el acuerdo de evaluación, la implementación y el tiempo docente. Se planteó una diferencia positiva inmediata como hipótesis principal; la retención y los recursos se interpretaron atendiendo a su incertidumbre y a los límites del diseño no aleatorizado.

Materiales y métodos

Diseño y contexto del caso

Se analizó una base organizada como un cuasiexperimento prospectivo con dos condiciones, dos mediciones previas y dos posteriores. El escenario comprendió 480 registros de estudiantes mayores de edad, 24 paralelos de 20 integrantes y 12 pares, distribuidos en tres instituciones identificadas como A, B y C. Cada institución incluyó cuatro pares. Se representaron asignaturas universitarias con un componente de escritura académica en español; cada par compartió asignatura, institución y un mismo perfil docente. Las condiciones se definieron a nivel de paralelo, con uno asistido y otro de comparación por par. La asignación se representó como administrativa y no se interpretó como aleatoria.

El calendario se situó entre el 15 de septiembre y el 8 de diciembre de 2025. Las mediciones basales se ubicaron el 15 y el 22 de septiembre; las seis unidades, entre el 29 de septiembre y el 7 de noviembre; la evaluación inmediata, el 10 de noviembre; y el seguimiento, el 8 de diciembre. Las fechas definieron el orden temporal del escenario. La unidad de análisis fue el registro individual y la unidad de asignación y agrupamiento fue el paralelo. Se utilizaron los componentes pertinentes de TREND para explicitar la evaluación no aleatorizada (Des Jarlais et al., 2004) y TIDieR para describir la intervención (Hoffmann et al., 2014). No se declaró reclutamiento, pilotaje de campo o registro público de un ensayo.

La justificación inicial consideró una diferencia estandarizada de 0,35, a bilateral de 0,05 y potencia del 90 %. La comparación no ajustada requirió aproximadamente 344 observaciones independientes. Al incorporar una correlación basal-final de 0,60, un efecto de diseño de 1 + 19 × 0,03 = 1,57 y un 15 % de pérdidas, la aproximación 344 × (1 − 0,60²) × 1,57 / 0,85 produjo 407 registros iniciales. Se utilizó un tamaño de 480 para representar 24 paralelos equilibrados. Esta aproximación se complementó con 2000 réplicas de un diseño con 12 pares, 17 desenlaces por paralelo y la misma regla ANCOVA-CR2. La potencia fue del 90,9 % con correlación intraclase residual de 0,03 y del 81,2 % con 0,06; los errores de Monte Carlo fueron de 0,64 y 0,87 puntos porcentuales. Se mantuvieron explícitos los supuestos, siguiendo Lakens (2022). La evaluación de potencia fue una comprobación del diseño y no potencia calculada a partir del efecto obtenido.

Estimando y justificación del tamaño

El estimando principal fue la diferencia media ajustada entre condiciones en la puntuación inmediata de un texto nuevo elaborado sin asistencia, conservando todos los registros en su condición original. Este contraste incluyó los distintos niveles de participación representados en la base. Su interpretación causal hubiera requerido ausencia de confusión residual y comparabilidad entre condiciones; el emparejamiento y el ajuste no garantizaron esos supuestos. El resultado del seguimiento se definió como secundario jerárquico y los indicadores de implementación y recursos como exploratorios.

Intervención y comparación activa

Se configuró una secuencia de seis unidades semanales, con dos sesiones de 60 minutos por semana: 12 sesiones y 12 horas de trabajo programado por registro. Ambas condiciones compartieron dossier, objetivos, rúbrica, dos ciclos de comentarios por unidad y una introducción común a integridad académica y verificación de fuentes. La condición asistida representó un flujo en el que la IA produjo la primera versión de comentarios y el docente verificó su pertinencia antes de entregarlos. En la comparación, el docente elaboró los comentarios sin generación automática y utilizó los mismos criterios y límite de extensión. El tiempo estudiantil se mantuvo equivalente en el diseño; el tiempo docente se trató como resultado operativo.

Tabla 1.

Secuencia de seis unidades de escritura

Semana	Objetivo	Actividad y evidencia de revisión
1	Delimitar una tesis	Formulación de una posición y justificación de su alcance.
2	Vincular evidencia y afirmaciones	Selección de fuentes del dossier y explicación de su pertinencia.
3	Examinar contraargumentos	Construcción de una objeción plausible y respuesta razonada.
4	Mejorar la coherencia	Revisión de la secuencia argumental y de los vínculos entre ideas.
5	Verificar atribución y exactitud	Comprobación de citas, paráfrasis y correspondencia documental.
6	Integrar el argumento	Texto completo, registro de decisiones y versión final revisada.

Fuente: Elaboración propia.

 https://doi.org/10.5281/zenodo.23354388

La instrucción operativa se definió del siguiente modo: «Analiza el borrador únicamente con el dossier y la rúbrica. Identifica como máximo dos problemas prioritarios. Para cada uno, señala el fragmento pertinente, formula una pregunta que ayude a revisarlo y explica el criterio relacionado. No escribas una versión sustitutiva ni añadas hechos o fuentes ajenos al dossier. Indica las afirmaciones que no puedas verificar. Limita la respuesta a 200 palabras». Se especificó la aceptación, modificación o sustitución de cada comentario antes de su entrega. Esta instrucción describió el componente pedagógico del caso; no se ejecutó un proveedor ni se atribuyeron los resultados a una marca, versión o configuración comercial de IA.

Evaluación de escritura y rúbrica

Las cuatro evaluaciones representaron tareas nuevas de 60 minutos, sin acceso a IA y con dossier de consulta. Se utilizaron cuatro formas, rotadas de modo equilibrado: cada paralelo contó con cinco registros por forma en cada momento y cada registro recorrió las cuatro formas. El equilibrio controló su distribución; no constituyó evidencia de equivalencia lingüística o de dificultad. El desenlace principal se obtuvo sumando cinco criterios de 0 a 4 puntos, con rango total de 0 a 20. La tabla 2 presenta los anclajes de puntuación. Los niveles 1 y 3 se definieron como desempeños intermedios entre los anclajes contiguos y se aplicaron de forma uniforme.

Tabla 2.

Rúbrica analítica de escritura argumentativa

Criterio	0 puntos	2 puntos	4 puntos
Tesis	No se identifica una posición.	La posición es reconocible, pero amplia o inestable.	La posición es precisa, sostenida y delimitada.
Evidencia	Las afirmaciones carecen de apoyo verificable.	Las fuentes son pertinentes, con conexión parcial al argumento.	La evidencia se integra y se reconocen sus límites.
Contraargumentación	No se consideran objeciones.	Se presenta una objeción con respuesta incompleta.	Se examina una objeción relevante y se responde razonadamente.
Coherencia	Las ideas aparecen inconexas.	La organización se comprende, con saltos de razonamiento.	La secuencia es lógica y explicita relaciones entre ideas.
Atribución y exactitud	Hay fuentes inventadas o atribución gravemente incorrecta.	La atribución es parcial, con errores localizados.	Las citas y paráfrasis son exactas, recuperables y pertinentes.

Nota. Los cinco criterios tuvieron la misma ponderación. La rúbrica delimitó el desempeño escrito; no constituyó una medida de pensamiento crítico general ni una herramienta para atribuir infracciones académicas.

Se generaron dos calificaciones independientes por texto y se utilizó la media de sus totales. Las discrepancias superiores a cuatro puntos se resolvieron mediante una tercera calificación y la mediana de los tres totales. El acuerdo se calculó antes de esa resolución mediante ICC de acuerdo absoluto para la media de dos evaluadores, ICC(A,2), con un modelo de dos vías y 1000 remuestreos de pares. La elección de la forma de ICC siguió la distinción metodológica de Koo y Li (2016). El diseño de aplicación contempló evaluación anónima y orden mezclado de textos; al no existir ensayos de personas, no se afirmó que jueces reales hubieran aplicado un procedimiento de cegamiento o validación.

Construcción de la base y registros operativos

Los registros se construyeron con la semilla 20261011. El mecanismo incorporó variación por institución a través de los pares, por paralelo y por individuo; dos puntuaciones basales correlacionadas; una diferencia inicial de composición entre condiciones; y componentes de cambio posterior. Se fijaron diferencias generadoras de 1,20 puntos en la medición inmediata y 0,75 en el seguimiento sobre la variable continua subyacente. Las puntuaciones de rúbrica se obtuvieron al generar cinco criterios, añadir error de evaluación, redondear a enteros entre 0 y 4 y sumar. Los valores generadores no se interpretaron como parámetros conocidos de una población. El código conservó las evaluaciones originales y las decisiones de resolución de discrepancias.

Se establecieron 38 evaluaciones inmediatas ausentes y 72 de seguimiento, con un patrón monótono: quienes carecieron de evaluación inmediata también carecieron de seguimiento. La selección dependió del nivel basal, las horas de trabajo y la primera generación universitaria, manteniendo completas las covariables iniciales. El mecanismo fue compatible con ausencia al azar condicionada a información observada. También se generaron indicadores de asistencia, ciclos completados, uso externo de IA, errores, modificaciones y tiempo docente. Estos indicadores representaron bitácoras numéricas; no procedieron de respuestas de un sistema, entrevistas o auditorías de comentarios reales.

La adherencia se definió como al menos 10 de las 12 sesiones; se describió sin excluir registros del análisis principal. El tiempo docente por ciclo incluyó elaboración o generación, verificación, modificación y entrega. Un ciclo con rechazo incluyó la sustitución del comentario y no se contabilizó como un comentario entregado adicional. Se registró el uso de IA externa al flujo previsto en ambas condiciones, sin usar detectores automáticos para atribuir autoría. No se estimaron costos de licencia, infraestructura o capacitación, de modo que el tiempo representado no equivalió a una evaluación económica completa.

Análisis estadístico

Se describieron las variables iniciales mediante medias, desviaciones estándar y frecuencias. La comparabilidad se examinó con diferencias estandarizadas entre condiciones, sin utilizar significación basal como criterio de selección de covariables. La diferencia entre los cambios de las dos mediciones previas se estimó con ajuste por par y corrección de agrupamiento. Este contraste se utilizó como diagnóstico y no como demostración de tendencias paralelas. No se eliminaron registros por valores de desempeño ni se perdieron paralelos completos.

El análisis principal aplicó ANCOVA a la puntuación inmediata, con condición, media de las dos puntuaciones basales, indicadores de forma de evaluación y efectos fijos de par. Se estimaron errores estándar robustos CR2 agrupados por paralelo y grados de libertad de Satterthwaite, apropiados para tratar la incertidumbre con pocos conglomerados (Pustejovsky y Tipton, 2018). La institución y el perfil docente quedaron absorbidos por los efectos de par. La diferencia estandarizada se calculó dividiendo la diferencia ajustada por la desviación basal combinada; no se denominó g de Hedges ni se calculó usando la desviación posterior. Sus intervalos se escalaron con el mismo denominador basal, tratado como fijo.

Se conservaron los 480 registros mediante 100 imputaciones múltiples secuenciales. Primero se imputó la evaluación inmediata a partir del promedio y cambio basal, edad, género, primera generación, trabajo, experiencia previa con IA, forma e indicadores de paralelo. Después se imputó el seguimiento incorporando esas variables, su forma y la puntuación inmediata, disponible o imputada. Los indicadores de paralelo representaron condición y par sin introducir columnas redundantes. Se realizaron extracciones de varianza de una distribución inversa chi cuadrado escalada y de coeficientes de su distribución normal condicional, seguidas de error individual; los valores se acotaron a 0–20. Se combinaron estimaciones y varianzas CR2 con las reglas de Rubin y grados de libertad ajustados de Barnard y Rubin (1999). Esta combinación se interpretó como aproximación, sin afirmar cobertura exacta.

Se utilizó una secuencia jerárquica: el contraste del seguimiento se evaluó a $\alpha = 0{,}05$ únicamente tras rechazar la hipótesis principal. Los demás análisis fueron exploratorios. La diferencia entre seguimiento e inmediato se analizó como un cambio por registro, preservando la dependencia entre mediciones; esta parametrización sustituyó un modelo longitudinal más amplio. La sensibilidad incluyó casos completos, ajuste adicional por cambio basal y covariables socioeducativas, diferencias en diferencias respecto de la segunda medición basal y exclusión sucesiva de cada par. El contraste de cambio utilizó la forma de seguimiento, que determinó la forma inmediata por la rotación establecida.

Para evaluar ausencia no al azar se redujeron en 1, 2 y 3 puntos las puntuaciones imputadas de intervención y se aumentaron en la misma cantidad las de comparación, conservando los límites 0–20. Estos escenarios fueron desfavorables a la modalidad asistida y se aplicaron por separado a cada desenlace. Se informaron intervalos del 95 % y valores p bilaterales. El tiempo docente se resumió con mediana y rango intercuartílico; la diferencia de tiempos medios se estimó a partir de los 12 contrastes entre condiciones dentro de cada par, con distribución t y 11 grados de libertad. No se trató cada comentario como una observación experimental independiente. Los cálculos se realizaron con Python 3.12.14, NumPy 2.3.5, SciPy 1.17.0 y pandas 2.2.3.

Consideraciones éticas y reproducibilidad

El ejercicio no involucró participantes, textos identificables ni datos enviados a un proveedor. No se atribuyó aprobación ética, consentimiento, autorización institucional o implementación a personas o universidades reales. La base, las calificaciones, las bitácoras y el código se conservaron en un anexo reproducible. Para una aplicación de campo, la identidad del modelo, sus versiones, los materiales completos y las salidas efectivamente recibidas serían componentes necesarios de la intervención; su

ausencia aquí delimitó el alcance del ejemplo y evitó presentar el caso como validación de un producto comercial.

Resultados

Flujo y características iniciales

La base incluyó 240 registros por condición y 12 paralelos por grupo. En la evaluación inmediata se dispuso de 220 registros de intervención y 222 de comparación, correspondientes al 92,1 % del total. En el seguimiento se conservaron 206 y 202, respectivamente, equivalentes al 85,0 %. No hubo cambios de condición ni pérdida de paralelos completos. El menor número de evaluaciones disponibles por paralelo fue de 16 en el momento inmediato y 14 en el seguimiento. Las ausencias se trataron como datos no disponibles, sin asignarles motivos administrativos o personales que no estuvieron representados.

Tabla 3.

Características basales por condición

Variable	IA y revisión docente	Retroalimentación docente	Diferencia estandarizada
Edad en años	22,10 ± 2,73	22,09 ± 2,75	0,006
Mujeres	140 (58,3 %)	128 (53,3 %)	0,101
Primera generación universitaria	90 (37,5 %)	106 (44,2 %)	-0,136
Trabajo remunerado en horas semanales	5,44 ± 8,03	4,73 ± 8,55	0,086
Experiencia previa con IA	116 (48,3 %)	95 (39,6 %)	0,177
Primera puntuación basal	10,02 ± 2,40	10,22 ± 2,56	-0,082
Segunda puntuación basal	10,26 ± 2,20	10,34 ± 2,57	-0,034
Promedio de ambas puntuaciones basales	10,14 ± 2,18	10,28 ± 2,44	-0,061

Nota. n = 240 por condición. Variables continuas: media ± desviación estándar; categóricas: n (%). La diferencia estandarizada se calculó como intervención menos comparación. La primera generación indicó ausencia de titulación universitaria en ambos progenitores.

Las diferencias estandarizadas absolutas de las variables iniciales presentadas no superaron 0,177. El promedio basal fue ligeramente menor en intervención, con una diferencia estandarizada de −0,061. La diferencia entre cambios previos fue de 0,12 puntos (IC del 95 %: −0,14 a 0,39; p = 0,336). Este resultado no mostró una incompatibilidad clara entre trayectorias iniciales, pero dos mediciones y un intervalo que incluyó el cero no demostraron intercambiabilidad ni descartaron confusión.

Puntuaciones y acuerdo entre evaluaciones

Tabla 4.

Puntuaciones de escritura de 0 a 20 en cada momento

Momento	IA y revisión docente	Retroalimentación docente
Basal 1	240; 10,02 ± 2,40	240; 10,22 ± 2,56
Basal 2	240; 10,26 ± 2,20	240; 10,34 ± 2,57
Inmediato	220; 12,65 ± 2,40	222; 11,18 ± 2,50
Seguimiento a cuatro semanas	206; 12,08 ± 2,22	202; 11,30 ± 2,31

Nota. Cada celda presenta n; media ± desviación estándar. Las medias corresponden a evaluaciones disponibles; no son medias ajustadas ni combinadas entre imputaciones.

Los ICC(A,2) fueron 0,931 y 0,923 en las dos evaluaciones iniciales, 0,929 en la inmediata (IC del 95 %: 0,916 a 0,938) y 0,915 en el seguimiento (IC del 95 %: 0,904 a 0,923). Se identificaron dos discrepancias superiores a cuatro puntos en el momento inmediato y se aplicó la regla de tercera evaluación. El acuerdo describió la coherencia de las calificaciones construidas; no sustituyó evidencia de validez de contenido, equivalencia de formas o funcionamiento de la rúbrica en una población real.

Diferencias ajustadas entre condiciones

Tabla 5.

Diferencias ajustadas en escritura autónoma

Momento	Análisis	Diferencia	IC 95 %	p
Inmediato	Principal con imputación	1,61	0,96 a 2,25	< 0,001
Inmediato	Casos completos	1,60	0,99 a 2,22	< 0,001
Inmediato	Ajuste adicional	1,58	0,92 a 2,23	< 0,001
Cuatro semanas	Principal con imputación	0,95	0,58 a 1,32	< 0,001
Cuatro semanas	Casos completos	0,95	0,69 a 1,21	< 0,001
Cuatro semanas	Ajuste adicional	0,95	0,58 a 1,32	< 0,001

Nota. Diferencia en puntos, intervención menos comparación. Modelo principal: promedio basal, forma y efectos de par; errores CR2 por paralelo. Imputación: n = 480. Casos completos: n = 442 y 408. El ajuste adicional incluyó cambio basal, edad, género, primera generación, trabajo y experiencia previa con IA.

La diferencia inmediata ajustada fue de 1,61 puntos a favor de la modalidad asistida (IC del 95 %: 0,96 a 2,25; p < 0,001). Al utilizar la desviación basal combinada de 2,31, la diferencia estandarizada fue de 0,70 (IC del

95 %: 0,42 a 0,97). La estimación se mantuvo cercana en casos completos y con covariables adicionales. En el seguimiento, la diferencia fue de 0,95 puntos (IC del 95 %: 0,58 a 1,32; p < 0,001), equivalente a 0,41 desviaciones basales. El contraste del seguimiento cumplió la condición jerárquica definida para su interpretación.

La reducción estimada de la diferencia entre condiciones desde el momento inmediato al seguimiento fue de 0,66 puntos, pero el intervalo del contraste de cambio incluyó el cero: −0,66 (IC del 95 %: −1,34 a 0,02; p = 0,055). Por tanto, la menor estimación puntual del seguimiento no bastó para demostrar una disminución de la ventaja. Los grados de libertad efectivos de los análisis principales fueron 9,20 y 7,44; estos valores reflejaron la información aportada por los conglomerados y las imputaciones, no el número total de registros.

Sensibilidad a datos faltantes y especificación

La diferencia inmediata permaneció positiva en los tres escenarios adversos. En cambio, el seguimiento mostró mayor dependencia del supuesto de ausencia: con un desplazamiento de dos puntos por condición, la estimación fue de 0,35 (IC del 95 %: −0,02 a 0,71; p = 0,060), y con tres puntos fue de 0,04 (IC del 95 %: −0,33 a 0,42). Esto identificó un límite concreto a la robustez de la retención y no probó que algunode esos mecanismos fuera el verdadero.

Tabla 6.

Escenarios desfavorables para las evaluaciones ausentes

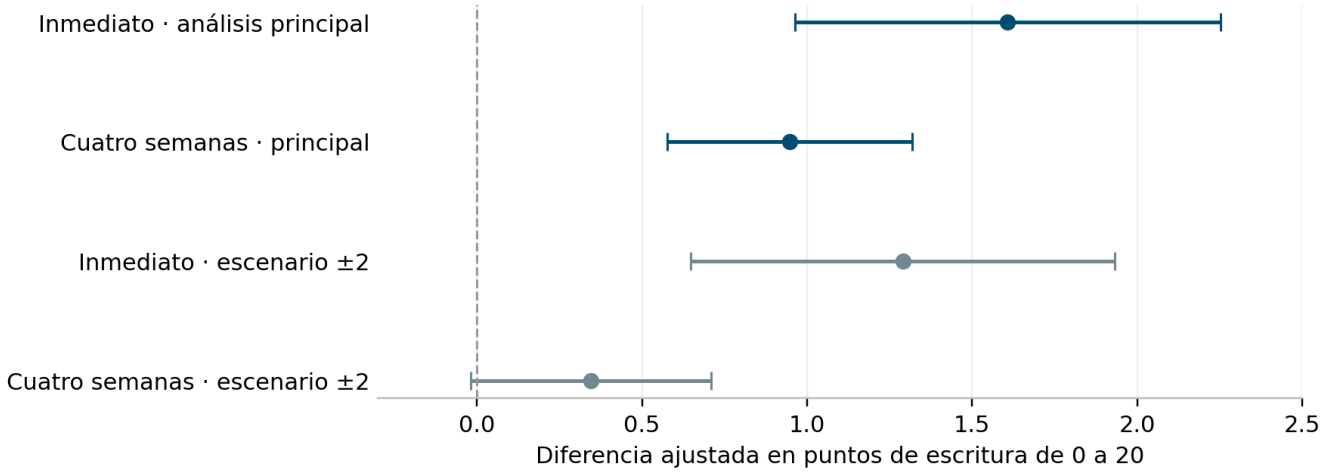
Escenario	Diferencia inmediata [IC del 95 %]	Diferencia a cuatro semanas [IC del 95 %]
Ausencia al azar	1,61 [0,96 a 2,25]	0,95 [0,58 a 1,32]
Desplazamiento adverso de 1 punto(s)	1,45 [0,81 a 2,09]	0,65 [0,28 a 1,01]
Desplazamiento adverso de 2 punto(s)	1,29 [0,65 a 1,93]	0,35 [-0,02 a 0,71]
Desplazamiento adverso de 3 punto(s)	1,13 [0,48 a 1,78]	0,04 [-0,33 a 0,42]

Nota. Cada desplazamiento redujo las notas imputadas de intervención y aumentó en igual cantidad las de comparación. No se modificaron evaluaciones disponibles. Los escenarios se aplicaron por separado a cada momento.

Las diferencias en diferencias fueron de 1,58 puntos inmediatamente (IC del 95 %: 0,91 a 2,25) y 0,93 a cuatro semanas (IC del 95 %: 0,48 a 1,38). Al excluir sucesivamente cada par en casos completos, las estimaciones oscilaron entre 1,50 y 1,78 para el resultado inmediato, y entre 0,88 y 1,00 para el seguimiento. El error de Monte Carlo del coeficiente principal fue de 0,0045 y 0,0072, respectivamente, equivalente al 1,6 % y 4,6 % de su error estándar. Solo una de las 11000 puntuaciones imputadas necesitó acotarse al intervalo 0–20.

Figura 1.

Diferencias ajustadas y sensibilidad a las evaluaciones ausentes



Nota. Barras: IC del 95 %. Valores positivos: mayor puntuación en intervención. El escenario ±2 redujo en dos puntos las puntuaciones imputadas de intervención y aumentó en dos las de comparación.

Implementación representada y tiempo docente

Los 2738 ciclos de intervención y los 2676 de comparación incluyeron revisión docente en las bitácoras. En intervención se registraron 223 indicadores de error inicial y 57 rechazos con sustitución. Tras la revisión, los indicadores de error residual fueron ocho y siete, respectivamente; su baja frecuencia no permitió concluir equivalencia de calidad. La mediana

de tiempo fue de 4,96 minutos en intervención y 7,23 en comparación. La diferencia media entre condiciones dentro de los 12 pares fue de −2,32 minutos por ciclo (IC del 95 %: −2,44 a −2,19). Los tiempos acumulados fueron 231,88 y 329,92 horas, pero sus diferentes números de ciclos y la omisión de costos de implementación impidieron interpretarlos como ahorro presupuestario neto.

Tabla 7.

Indicadores de implementación y recursos

Indicador	IA y revisión docente	Retroalimentación docente
Asistencias a sesiones programadas	2583/2880 (89,7 %)	2564/2880 (89,0 %)
Registros con al menos diez sesiones	210/240 (87,5 %)	201/240 (83,8 %)
Ciclos de comentarios completados	2738/2880 (95,1 %)	2676/2880 (92,9 %)
Comentarios con error inicial	223/2738 (8,1 %)	77/2676 (2,9 %)
Comentarios modificados o sustituidos	776/2738 (28,3 %)	325/2676 (12,1 %)
Comentarios con indicador de error residual	8/2738 (0,3 %)	7/2676 (0,3 %)
Registros con uso externo de IA	23/240 (9,6 %)	20/240 (8,3 %)
Tiempo docente por ciclo en minutos	4,96 [4,21–5,80]	7,23 [6,12–8,39]

Nota. El tiempo se expresó como mediana [percentiles 25–75]. Los indicadores procedieron de bitácoras numéricas del caso. Los comentarios sustituidos se incluyeron entre los modificados y no aumentaron el denominador de ciclos.

Discusión

El análisis ilustró una ventaja ajustada de 1,61 puntos en escritura autónoma inmediatamente después de la secuencia de retroalimentación y una diferencia de 0,95 a cuatro semanas. La estabilidad del resultado inmediato frente a ajustes y escenarios de ausencia fue mayor que la del seguimiento. La retención dejó de presentar un intervalo completamente positivo cuando las puntuaciones ausentes se desplazaron de forma adversa en dos puntos por condición. Esta distinción impidió resumir todos los resultados como una mejora igualmente robusta. Además, la comparación perteneció a un escenario construido con diferencias generadoras y no estimó la eficacia de un sistema utilizado por estudiantes reales.

La elección de un texto nuevo sin asistencia respondió a la distinción entre aprendizaje y desempeño. Soderstrom y Bjork (2015) explicaron que un buen resultado durante la práctica puede no reflejar cambios duraderos en las capacidades. Stevenson y Phakiti (2014) ya habían señalado esta dificultad en la literatura sobre evaluación automática de escritura. Una intervención puede mejorar la versión que se entrega porque facilita correcciones puntuales, sin desarrollar los criterios que permiten revisar otra tarea. Por ello, el desenlace autónomo resultó más pertinente que la calidad del borrador asistido para la pregunta formulada, aunque una prueba cercana en contenido solo representó transferencia próxima.

La comparación con estudios de IA requirió diferenciar sus desenlaces y controles. Steiss et al. (2024) analizaron la calidad de comentarios humanos y automáticos; no estimaron la transferencia posterior del aprendizaje. Meyer et al. (2024) compararon retroalimentación generativa con revisión sin comentarios y midieron la mejora del texto revisado. En cambio, el presente esquema utilizó retroalimentación docente como comparación activa. Estas diferencias impidieron ordenar los estudios como si todos midieran el mismo efecto. El metaanálisis de Fleckenstein et al. (2023) proporcionó una base para estudiar automatización, pero su efecto promedio tampoco constituyó un estándar de eficacia para esta modalidad concreta.

La supervisión docente se consideró parte constitutiva de la intervención. Las revisiones de Hattie y Timperley (2007), Shute (2008) y Wisniewski et al. (2020) respaldan valorar la información que orienta acciones sobre la tarea. La generación de comentarios solo ofreció una primera versión dentro del flujo representado; su verificación, priorización y eventual sustitución modificaron el producto que habría recibido el estudiante. En consecuencia, aun con datos de campo, una diferencia entre condiciones correspondería al paquete de trabajo y no demostraría que el componente generativo aislado fuera superior al juicio docente.

La importancia del diseño del apoyo también aparece en dominios distintos de la escritura. Kestin et al. (2025) evaluaron una tutoría pedagógicamente estructurada en física, mientras que Bastani et al. (2025) distinguieron entre acceso generativo sin restricciones y apoyos con salvaguardas en matemáticas. Ninguno de esos trabajos permite anticipar una magnitud para escritura argumentativa universitaria en español. Su aportación para este

caso fue fundamentar preguntas orientadoras, evitar la sustitución del trabajo intelectual y evaluar después sin acceso a la herramienta. La coincidencia de una dirección favorable no constituye replicación cuando cambian población, contenido y modalidad.

Las decisiones del estudiante siguen siendo el vínculo entre comentarios y revisión. Boud y Molloy (2013) propusieron organizar la retroalimentación de modo que los aprendices desarrollen capacidades para juzgar y dirigir su aprendizaje. Carless y Boud (2018) y Winstone et al. (2017) precisaron componentes de esa participación. Una explicación útil puede ser ignorada, malinterpretada o aceptada sin comprensión; por eso, contabilizar comentarios no establece que el proceso formativo ocurriera. Los registros de cambios y razones deberían examinarse en una aplicación real, sin convertir asociaciones de proceso en una prueba de mediación causal.

El trabajo con rúbricas y autoevaluación ofreció una vía de aplicación coherente con esta perspectiva. Panadero et al. (2017) encontraron beneficios de la autoevaluación para procesos autorregulatorios, y Huisman et al. (2019) mostraron el potencial de la retroalimentación entre pares en escritura universitaria. En el curso, esas prácticas podrían permitir que el estudiante contraste una sugerencia generativa con criterios, ejemplos y observaciones de otros lectores. No se derivó de estos antecedentes que toda corrección deba aceptarse. La capacidad de explicar por qué se mantiene una formulación puede ser tan informativa como la cantidad de cambios realizados.

El tiempo docente constituyó una dimensión operativa separada del aprendizaje. Noy y Zhang (2023) demostraron que la asistencia generativa puede modificar la productividad en tareas profesionales, pero no evaluaron el tiempo total de un flujo universitario de retroalimentación con verificación. En las bitácoras del caso, la diferencia media fue de 2,32 minutos por ciclo, favorable a la modalidad asistida. Esa magnitud dependió de supuestos de generación y no incluyó preparación de materiales, capacitación, infraestructura o licencia. Una evaluación institucional posterior necesitaría registrar esos componentes, las incidencias técnicas y la distribución del tiempo, además de la mediana por comentario.

Los indicadores de error impidieron identificar rapidez con calidad. La intervención presentó más errores iniciales que la comparación y necesitó modificaciones y sustituciones; los errores residuales representados fueron poco frecuentes en ambos grupos. Esto no permitió asegurar equivalencia ni certificar que una herramienta fuera segura o exacta. Memarian y Doleck (2023) y Yan et al. (2024) muestran que responsabilidad y trazabilidad forman parte de la implementación educativa. En términos operativos, ello exige identificar quién revisa, qué considera un error, qué conserva en el registro y cómo puede el estudiante solicitar una explicación o cuestionar una recomendación.

Dawson et al. (2019) documentaron diferencias entre las perspectivas de docentes y estudiantes sobre qué hace eficaz la retroalimentación. Esta observación sugiere complementar las medidas de desempeño con evidencia de comprensión y uso de los comentarios, sin sustituir el desenlace principal por satisfacción. Kluger y DeNisi (1996) mostraron que las intervenciones de retroalimentación no producen beneficios uniformes. En el caso analizado, limitar el mensaje a dos problemas prioritarios pretendió mantener su foco, pero esa decisión pedagógica necesitaría contraste empírico. Ni la longitud del comentario ni la presencia de un tono amable demuestran por sí solas utilidad para la revisión.

El seguimiento requirió una interpretación particularmente prudente. Su estimación puntual fue menor que la inmediata, pero el contraste directo de reducción fue impreciso y no justificó afirmar un deterioro demostrado. A la vez, el escenario adverso de dos puntos reveló que la conclusión sobre retención dependió de los valores posibles de las evaluaciones ausentes. Ambas observaciones pueden coexistir: la comparación entre momentos y la sensibilidad a ausencia responden a preguntas distintas. La imputación permitió incorporar registros incompletos bajo supuestos explícitos; no recuperó información que hubiera sido observada ni identificó el mecanismo de pérdida.

El uso de 24 paralelos y 12 pares condicionó la inferencia. La corrección CR2 y los grados de libertad ajustados atendieron a ese número limitado de conglomerados, siguiendo Pustejovsky y Tipton (2018). La combinación con imputaciones reconoció incertidumbre adicional conforme a Barnard y Rubin (1999). Sin embargo, errores estándar adecuados no eliminan sesgos de asignación, diferencias de

expectativas o contaminación entre grupos. Los efectos de par controlaron características compartidas, pero una asignación administrativa podría relacionarse con factores no medidos que también influyan en la escritura. Las dos mediciones basales y la estabilidad al excluir pares fueron diagnósticos, no pruebas de ausencia de confusión.

Las limitaciones de medición fueron igualmente sustantivas. Un ICC elevado indicó acuerdo de las calificaciones representadas y no validez del constructo, distinción congruente con Koo y Li (2016). Las formas se distribuyeron de manera equilibrada, pero no se aplicó un pilotaje de equivalencia con textos reales. Tampoco se comprobó invariancia de criterios entre disciplinas o perfiles lingüísticos. El resultado total no permitió atribuir la diferencia a tesis, evidencia o contraargumentación por separado; hacerlo habría requerido análisis adicionales y atención a comparaciones múltiples. El escenario mantuvo grupos del mismo tamaño y no reprodujo todas las irregularidades de una intervención institucional.

El aporte metodológico consistió en integrar comparación activa, temporalidad, evaluación autónoma, incertidumbre por agrupamiento y análisis de ausencia dentro de un mismo reporte. TREND y TIDieR ayudaron a describir esas decisiones, pero su uso no convirtió un ejemplo docente en evidencia de campo (Des Jarlais et al., 2004; Hoffmann et al., 2014). Para una investigación posterior, las prioridades serían disponer de materiales y versión del sistema documentados, validar la evaluación, ampliar la diversidad de paralelos y utilizar asignación aleatoria por conglomerados cuando fuera viable. Un seguimiento más largo y tareas de mayor distancia respecto del entrenamiento permitirían examinar la permanencia y el alcance de la transferencia.

Conclusiones

En el caso analizado, la modalidad de retroalimentación apoyada por IA y revisada por docentes presentó una diferencia inmediata ajustada de 1,61 puntos en una escala de escritura de 0 a 20. La diferencia a cuatro semanas fue de 0,95 puntos y resultó más sensible a los supuestos sobre evaluaciones ausentes. La menor estimación del seguimiento no demostró por sí sola una reducción temporal de la ventaja. Los indicadores operativos representaron menor tiempo por ciclo y necesidad de verificación y corrección de comentarios. El ejercicio mostró cómo separar desempeño autónomo, retención, calidad y recursos dentro de un diseño cuasiexperimental. No estableció causalidad, eficacia de una plataforma ni resultados generalizables a estudiantes reales.

Declaraciones editoriales

Disponibilidad de datos

Los datos y materiales que respaldan los resultados de este estudio se encuentran incluidos en el artículo y, cuando corresponde, en sus archivos suplementarios. Las solicitudes de información adicional podrán dirigirse al autor de correspondencia y estarán sujetas a los principios de confidencialidad, protección de datos y ética de la investigación.

Declaración de ética

La investigación se desarrolló de conformidad con la legislación nacional vigente, las disposiciones institucionales y los principios éticos aplicables a estudios con participación humana. Debido a su naturaleza social y no experimental, sin intervención clínica, biomédica ni manipulación de variables, no requirió la aprobación de un comité de ética. Todos los participantes recibieron información suficiente sobre los objetivos y procedimientos del estudio y otorgaron su consentimiento informado por escrito. Asimismo, se garantizó la confidencialidad, el anonimato y el uso exclusivamente académico de la información recopilada.

Contribuciones de autoría

Las contribuciones de autoría se establecieron de acuerdo con la taxonomía CRediT:

DPG: Conceptualización; metodología; redacción del borrador original; redacción, revisión y edición.

LMMB: Metodología; investigación; redacción del borrador original; redacción, revisión y edición.

JCPS: Investigación; redacción del borrador original; redacción, revisión y edición.

TRSS: Conceptualización; supervisión; administración del proyecto; redacción, revisión y edición.

Financiamiento

Los autores declaran que no recibieron financiamiento específico de organismos públicos, instituciones privadas ni entidades sin fines de lucro para la realización o publicación de este estudio.

Conflicto de intereses

Los autores declaran que la investigación se desarrolló en ausencia de relaciones comerciales, financieras, profesionales o personales que pudieran influir de manera inapropiada en sus resultados, interpretación o publicación.

Declaración sobre el uso de inteligencia artificial generativa

Los autores declaran que no utilizaron herramientas de inteligencia artificial generativa para producir el contenido científico, analizar o interpretar los datos, formular los resultados o elaborar las conclusiones del manuscrito. La responsabilidad sobre la originalidad, exactitud, integridad y validez académica del contenido corresponde exclusivamente a los autores.

Responsabilidad editorial

Las afirmaciones, interpretaciones y conclusiones expresadas en el artículo son responsabilidad exclusiva de sus autores y no representan necesariamente la posición institucional de ARTI EDITORIAL S.A.S., de la revista, de sus editores o de las entidades a las que se encuentran afiliados los autores.

Referencias bibliográficas

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*(4), 948–955. https://doi.org/10.1093/biomet/86.4.948

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education, 38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education, 44*(1), 25–36. https://doi.org/10.1080/02602938.2018.1467877

Des Jarlais, D. C., Lyles, C., Crepaz, N., & TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94*(3), 361–366. https://doi.org/10.2105/AJPH.94.3.361

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, 1162454. https://doi.org/10.3389/frai.2023.1162454

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D. G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S. E., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A. W., & Michie, S. (2014). Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide. *BMJ, 348*, g1687. https://doi.org/10.1136/bmj.g1687

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific*

Reports, 15, 17458. https://doi.org/10.1038/s41598-025-97652-6

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lakens, D. (2022). Sample size justification. *Collabra: Psychology, 8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in artificial intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence, 5*, 100152. https://doi.org/10.1016/j.caeai.2023.100152

Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence, 6*, 100199. https://doi.org/10.1016/j.caeai.2023.100199

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218. https://doi.org/10.1080/03075070600572090

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Pustejovsky, J. E., & Tipton, E. (2018). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672–683. https://doi.org/10.1080/07350015.2016.1247004

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science, 10*(2), 176–199. https://doi.org/10.1177/1745691615569000

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 of students' writing. *Learning and Instruction, 91*, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://www.sciencedirect.com/science/article/pii/S1075293513000512

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370