CIENCIA INTERDISCIPLINARIA INTERNACIONAL

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

DOI: https://doi.org/10.70577/cieninter.v4i3.43

Retroalimentación automatizada mediante modelos de lenguaje y producción escrita en estudiantes de educación primaria de Guaranda, Ecuador

Automated Feedback Through Language Models and Written Production in Primary Education Students from Guaranda, Ecuador

Hortencia Magdalena Bonilla Verdezoto[1]
Escuela de Educación Básica Manuel de Echeandía
hortencia.bonilla@docentes.educacion.ec
https://orcid.org/0009-0006-3361-7578
Guranda – Ecuador

Zoila América García García[2]
Escuela de Educación Básica "21 de Abril"
zoilaa.garcia@docentes.educacion.edu.ec
https://orcid.org/0009-0007-5116-7600
Riobamba – Ecuador

Lidia Ivonne Basantes Silva [3]
Escuela de Educación Básica "21 de Abril"
lidia.basantes@docentes.educacion.edu.ec
https://orcid.org/0009-0006-1302-3123
Riobamba – Ecuador

Ximena Graciela Carguachi Mancheno[4]
Escuela de Educación Básica "21 de Abril"
ximena.carguachi@docentes.educacion.edu.ec
https://orcid.org/0009-0009-5233-6781
Riobamba – Ecuador

Cómo citar:
Bonilla Verdezoto, H. M., García García, Z. A., Basantes Silva , L. I., & Carguachi Mancheno, X. G. (2026). Retroalimentación automatizada mediante modelos de lenguaje y producción escrita en estudiantes de educación primaria de Guaranda, Ecuador. *Ciencia Interdisciplinaria Internacional*, *4*(3), 70–81. https://doi.org/10.70577/cieninter.v4i3.43

Fecha de recepción: 2026-07-03

Fecha de aceptación: 2026-07-15

Fecha de publicación: 2026-08-05

CIENCIA INTERDISCIPLINARIA INTERNACIONAL

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

RESUMEN

La integración de modelos de lenguaje de gran escala en el ámbito educativo ha generado un debate creciente sobre su potencial para apoyar procesos de escritura en los primeros niveles de escolarización. El presente estudio analizó los efectos de la retroalimentación automatizada generada por un modelo de lenguaje sobre la calidad de la producción escrita de estudiantes de quinto y sexto grado de educación general básica en instituciones educativas fiscales de la ciudad de Guaranda, provincia de Bolívar, Ecuador. Se empleó un diseño cuasiexperimental con grupo de control y grupo experimental, mediciones pretest y postest, y una muestra de 127 estudiantes distribuidos en cuatro aulas de dos unidades educativas. El grupo experimental recibió retroalimentación generada por un modelo de lenguaje adaptado, mientras que el grupo de control continuó con retroalimentación exclusivamente docente. La producción escrita se evaluó mediante una rúbrica analítica validada que contempló cinco dimensiones: coherencia, cohesión, adecuación léxica, corrección gramatical y estructura textual. Los resultados mostraron diferencias estadísticamente significativas a favor del grupo experimental en coherencia ($p = .003$), cohesión ($p = .011$) y corrección gramatical ($p = .007$), con tamaños de efecto medianos según la d de Cohen. No se encontraron diferencias en adecuación léxica ni en estructura textual. Entrevistas semiestructuradas con 12 docentes revelaron percepciones mixtas: valoraron la inmediatez del feedback, pero expresaron preocupaciones sobre la dependencia tecnológica y la limitada sensibilidad contextual del modelo. Los hallazgos sugieren que la retroalimentación automatizada puede complementar, aunque no sustituir, la labor docente en la enseñanza de la escritura en educación primaria.

Palabras clave: retroalimentación automatizada, modelos de lenguaje, producción escrita, educación primaria, inteligencia artificial en educación, evaluación formativa.

ABSTRACT

The integration of large language models in education has sparked growing debate regarding their potential to support writing processes at the earliest levels of schooling. This study examined the effects of automated feedback generated by a language model on the quality of written production among fifth- and sixth-grade students in public schools in Guaranda, Bolívar Province, Ecuador. A quasi-experimental design was used with control and experimental groups, pretest and posttest measurements, and a sample of 127 students distributed across four classrooms from two schools. The experimental group received feedback generated by an adapted language model, while the control group continued with teacher-only feedback. Written production was assessed using a validated analytical rubric covering five dimensions: coherence, cohesion, lexical adequacy, grammatical accuracy, and textual structure. Results showed statistically significant differences favoring the experimental group in coherence ($p = .003$), cohesion ($p = .011$), and grammatical accuracy ($p = .007$), with medium effect sizes according to Cohen's d. No differences were found in lexical adequacy or textual structure. Semi-structured interviews with 12 teachers revealed mixed perceptions: they valued the immediacy of the feedback but expressed concerns about technological dependence and the model's limited contextual sensitivity. Findings suggest that automated feedback can complement, though not replace, the teacher's role in writing instruction at the primary level.

Keywords: automated feedback, language models, written production, primary education, artificial intelligence in education, formative assessment.

CIENCIA INTERDISCIPLINARIA INTERNACIONAL

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

INTRODUCCIÓN

La enseñanza de la escritura en la educación primaria enfrenta tensiones que no son nuevas, pero que la irrupción tecnológica reciente ha vuelto más visibles. Por un lado, los currículos nacionales de la mayoría de países latinoamericanos reconocen la producción textual como una competencia transversal que debe desarrollarse desde los primeros años de escolarización (Ministerio de Educación del Ecuador, 2021). Por otro lado, las condiciones reales del aula dificultan un acompañamiento individualizado: ratios elevados de estudiantes por docente, tiempo limitado para la corrección de textos y una formación inicial del profesorado que rara vez aborda estrategias de retroalimentación escrita con profundidad suficiente (Flotts et al., 2021).

En este escenario, los modelos de lenguaje de gran escala (large language models, LLM) han captado la atención de investigadores y responsables de política educativa. Desde la aparición pública de ChatGPT a finales de 2022, la literatura ha documentado un interés acelerado en las posibilidades de estos sistemas para generar retroalimentación sobre textos académicos en educación superior (Kasneci et al., 2023; Yan et al., 2024). Sin embargo, la producción científica referida a la educación primaria es considerablemente menor, y los estudios realizados en contextos latinoamericanos resultan todavía escasos (Crompton y Burke, 2023).

Guaranda, capital de la provincia de Bolívar en la sierra central del Ecuador, presenta características que hacen de ella un caso particularmente relevante para este tipo de indagaciones. Se trata de una ciudad de tamaño intermedio, con una población estudiantil heterogénea en términos socioeconómicos y culturales, donde coexisten hablantes de español como lengua materna y estudiantes con influencia del kichwa en su entorno familiar. Las evaluaciones nacionales Ser Estudiante aplicadas entre 2021 y 2024 han ubicado a la provincia de Bolívar en los percentiles inferiores de desempeño en lengua y literatura en educación general básica (Instituto Nacional de Evaluación Educativa [INEVAL], 2023). Comprender si una herramienta de retroalimentación automatizada puede incidir positivamente en las habilidades de escritura de esta población adquiere, por tanto, pertinencia tanto teórica como práctica.

La retroalimentación formativa ha sido reconocida durante décadas como uno de los factores con mayor incidencia en el aprendizaje (Hattie y Timperley, 2007). En el dominio específico de la escritura, los trabajos de Graham et al. (2022) han reiterado que la calidad, la oportunidad y la especificidad del feedback influyen de manera directa en la revisión de los textos y, a mediano plazo, en la transferencia de habilidades compositivas. Los sistemas automatizados de evaluación de escritura (automated writing evaluation, AWE) no son recientes: herramientas como e-rater y Criterion operan desde inicios de siglo (Warschauer y Grimes, 2008). Lo que distingue a la generación actual de herramientas basadas en LLM es su capacidad para generar comentarios en lenguaje natural, con un grado de personalización que los motores anteriores, basados en reglas y modelos estadísticos más simples, difícilmente alcanzaban (Dai et al., 2023).

A pesar de estas promesas, varios interrogantes permanecen abiertos. ¿Resulta eficaz la retroalimentación de un LLM cuando los destinatarios son niños de entre 9 y 12 años, cuyas competencias lectoras pueden limitar la comprensión del feedback recibido? ¿Qué dimensiones de la escritura se ven más favorecidas por este tipo de intervención? ¿Cómo perciben los docentes la incorporación de una herramienta que opera parcialmente fuera de su control pedagógico? La literatura disponible no ofrece respuestas concluyentes a estas preguntas cuando se sitúan en educación primaria y, menos aún, en el contexto socioeducativo andino.

Con base en estas consideraciones, el propósito de la presente investigación fue determinar los efectos de la retroalimentación automatizada generada por un modelo de lenguaje sobre la calidad de la producción escrita de estudiantes de quinto y sexto grado de educación general básica en dos instituciones educativas fiscales de Guaranda, Ecuador, durante el año lectivo 2024-2025. De manera complementaria, se recogieron las percepciones docentes respecto al uso de esta tecnología en su práctica pedagógica.

CIENCIA INTERDISCIPLINARIA INTERNACIONAL

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

REVISIÓN DE LITERATURA

Retroalimentación formativa y escritura en educación primaria

La retroalimentación formativa, entendida como la información que un agente proporciona al aprendiz acerca de su desempeño con la intención de reducir la distancia entre el estado actual y el esperado (Hattie y Timperley, 2007), constituye un eje articulador en la didáctica de la escritura. El modelo Feed Up-Feed Back-Feed Forward propuesto por estos mismos autores ha sido retomado en investigaciones recientes que enfatizan la necesidad de que el feedback no se limite a señalar errores, sino que ofrezca pistas para la revisión y orientaciones sobre el paso siguiente (Wisniewski et al., 2020).

En educación primaria, la retroalimentación escrita adopta formas particulares. Los trabajos de Graham et al. (2022) con muestras de escuelas primarias de lengua inglesa mostraron que los comentarios excesivamente generales (del tipo "muy bien" o "revisa tu ortografía") producen mejoras mínimas en borradores posteriores. En contraste, las observaciones que apuntan a segmentos concretos del texto y formulan preguntas orientadoras generan ciclos de revisión más productivos. Peterson y McClay (2021) documentaron que, en aulas de cuarto grado, los estudiantes que recibieron retroalimentación con indicaciones explícitas sobre coherencia entre oraciones duplicaron las revisiones sustantivas respecto al grupo que recibió solo marcas de error.

En el contexto hispanohablante, la investigación sobre retroalimentación escrita en primaria es más limitada. Sotomayor et al. (2021), en un estudio con profesores chilenos, encontraron que la mayor parte de las anotaciones docentes en cuadernos de estudiantes de tercero y cuarto básico se concentraba en aspectos superficiales del texto (ortografía, caligrafía) y dedicaba poca atención a la organización de las ideas o a la adecuación del registro. Flotts et al. (2021) confirmaron tendencias similares para la región andina a partir de datos del Tercer Estudio Regional Comparativo y Explicativo (TERCE) y su continuación.

Sistemas automatizados de evaluación de escritura: evolución y estado actual

Los sistemas de evaluación automatizada de escritura (AWE, por sus siglas en inglés) tienen una trayectoria de más de dos décadas. Las primeras herramientas, como Project Essay Grade (PEG) y e-rater, empleaban rasgos superficiales del texto (longitud de oraciones, diversidad léxica, frecuencia de errores) para asignar puntuaciones holísticas (Warschauer y Grimes, 2008). Estas herramientas resultaron útiles para evaluaciones a gran escala, pero su capacidad para ofrecer retroalimentación formativa individualizada era limitada.

La aparición de modelos basados en aprendizaje profundo y, particularmente, de arquitecturas transformer (Vaswani et al., 2017) modificó el panorama de forma sustancial. Los modelos de lenguaje preentrenados como GPT-4, LLaMA 2 y Gemini poseen capacidad para generar texto coherente, mantener interacciones conversacionales y producir comentarios específicos sobre fragmentos de texto delimitados (OpenAI, 2023). Dai et al. (2023) compararon la retroalimentación generada por GPT-4 con la producida por tutores humanos en 200 ensayos universitarios y encontraron un nivel de acuerdo inter-rater de .71 (kappa de Cohen) entre ambas fuentes, lo que indica una coincidencia sustancial, aunque no completa.

Otros trabajos han matizado estos resultados. Steiss et al. (2024) observaron que los LLM tienden a producir retroalimentación más elogiosa y menos directiva que los docentes humanos, y que esta tendencia se acentúa cuando el texto evaluado presenta deficiencias graves. Escalante et al. (2023) analizaron la retroalimentación de ChatGPT en textos argumentativos de estudiantes mexicanos de secundaria y reportaron que el modelo identificaba con precisión errores de concordancia y puntuación, pero tenía dificultades para evaluar la pertinencia de los argumentos en relación con el propósito comunicativo.

IA en educación primaria: evidencias y vacíos

La revisión sistemática de Crompton y Burke (2023), que analizó 138 estudios empíricos sobre IA en educación publicados entre 2016 y 2022, mostró que apenas el 11 % de las investigaciones se situaban en educación primaria y que la mayor parte de los trabajos se concentraba en matemáticas y ciencias, con escasa atención a las competencias lingüísticas. Una

ISSN: 3121-2948
Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

tendencia similar fue documentada por Zheng et al. (2023), quienes señalaron que la investigación sobre IA y aprendizaje de lenguas se ha centrado de manera desproporcionada en estudiantes universitarios y en la enseñanza del inglés como lengua extranjera.

En América Latina, la producción es aún más concentrada. Sánchez-Prieto et al. (2022) realizaron un mapeo de la investigación sobre IA en educación en Iberoamérica y constataron que Brasil, España y México acumulan el 78 % de las publicaciones, mientras que los países andinos (Ecuador, Perú, Bolivia, Colombia) representan menos del 9 %. Para el caso ecuatoriano, Valverde-Berrocoso et al. (2022) identificaron un crecimiento incipiente de estudios sobre tecnología educativa, pero con un predominio de enfoques descriptivos y una ausencia casi total de diseños experimentales o cuasiexperimentales que permitan establecer relaciones causales.

El vacío es, por tanto, triple: temático (pocos estudios sobre retroalimentación automatizada en escritura), poblacional (escasa atención a la educación primaria) y geográfico (subrepresentación del contexto andino y ecuatoriano). La presente investigación pretende contribuir a llenar esta intersección de ausencias.

Marco teórico: zona de desarrollo próximo y andamiaje digital

El diseño conceptual de esta investigación se apoyó en la noción de zona de desarrollo próximo (ZDP) formulada por Vygotsky (1978), entendida como la distancia entre lo que un aprendiz puede realizar de manera autónoma y lo que puede lograr con la guía de un agente más competente. La retroalimentación formativa opera, desde esta perspectiva, como un mecanismo de andamiaje que permite al estudiante reconocer la brecha entre su producción actual y los criterios de calidad esperados.

Belland (2014) amplió la noción de andamiaje al ámbito digital y distinguió entre andamiaje estático (ayudas fijas incorporadas en una herramienta) y dinámico (respuestas adaptativas a las acciones del aprendiz). Los LLM, en tanto generan retroalimentación específica para cada texto producido por el estudiante, se sitúan más cerca del polo dinámico, aunque con una diferencia respecto al andamiaje humano: carecen de acceso a la historia de aprendizaje del alumno, a su contexto emocional y a los acuerdos pedagógicos establecidos en el aula (Holmes et al., 2022). La tensión entre la adaptabilidad lingüística del modelo y su descontextualización pedagógica constituye una de las preguntas que este estudio aborda de manera empírica.

MATERIALES Y MÉTODOS

Diseño de investigación

Se adoptó un diseño cuasiexperimental con grupo de control no equivalente, mediciones pretest y postest, complementado con un componente cualitativo de alcance exploratorio. La elección de este diseño responde a dos razones: la imposibilidad de aleatorizar la asignación de estudiantes a grupos, dado que las unidades educativas participantes organizan sus aulas de forma preexistente, y la necesidad de contar con una medición inicial que permitiera controlar las diferencias basales entre grupos (Shadish et al., 2002). El componente cualitativo consistió en entrevistas semiestructuradas con docentes de los grupos participantes y se incorporó para capturar dimensiones de la experiencia pedagógica que los datos cuantitativos no recogen.

Contexto y participantes

La investigación se desarrolló en dos unidades educativas fiscales de la zona urbana de Guaranda, seleccionadas mediante muestreo intencional con base en tres criterios: acceso a conectividad a internet en el aula, disposición del equipo directivo para participar en el estudio y contar con al menos dos paralelos de quinto o sexto grado de educación general básica (EGB). Guaranda, con una población estimada de 25 000 habitantes en su casco urbano (INEC, 2022), es la capital de la provincia de Bolívar, ubicada en la sierra central ecuatoriana a 2 668 metros sobre el nivel del mar.

La muestra estuvo conformada por 127 estudiantes: 64 en el grupo experimental (dos aulas: una de quinto y una de sexto grado) y 63 en el grupo de control (dos aulas equivalentes en la segunda institución). La edad de los participantes osciló entre 9 y 12 años (M = 10.4, DE = 0.87). El 52 % de los estudiantes eran mujeres y el 48 %, varones. En cuanto al componente cualitativo, participaron 12 docentes (8 mujeres, 4 hombres) con

ISSN: 3121-2948
Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

una experiencia profesional media de 14 años (rango: 5 a 28 años).

Variables e instrumentos

La variable independiente fue el tipo de retroalimentación recibida por los estudiantes: automatizada, generada por un modelo de lenguaje (grupo experimental), o exclusivamente docente (grupo de control). La variable dependiente fue la calidad de la producción escrita, operacionalizada a través de cinco dimensiones evaluadas mediante una rúbrica analítica: coherencia, cohesión, adecuación léxica, corrección gramatical y estructura textual. Cada dimensión se puntuó en una escala de 1 a 4 puntos, lo que generó un rango teórico de 5 a 20 puntos por texto.

La rúbrica fue adaptada a partir del instrumento desarrollado por Sotomayor et al. (2021) para la evaluación de textos narrativos en español en educación primaria. La adaptación incluyó la revisión de los descriptores para adecuarlos al currículo ecuatoriano de lengua y literatura, y fue sometida a juicio de cinco expertos (tres en didáctica de la lengua, uno en medición educativa y uno en lingüística computacional). El índice de validez de contenido (IVC) global resultó de .89, y la concordancia intercodificadores, calculada mediante el coeficiente kappa de Fleiss con dos evaluadores entrenados sobre una muestra de 30 textos piloto, alcanzó un valor de .81.

Para la recogida de percepciones docentes se diseñó una guía de entrevista semiestructurada con 11 preguntas organizadas en tres ejes: experiencia previa con tecnología educativa, valoración de la retroalimentación automatizada observada durante la intervención, y condiciones percibidas para la sostenibilidad de la herramienta. La guía fue piloteada con tres docentes de una escuela no participante y ajustada en función de sus observaciones.

Herramienta de retroalimentación automatizada

Para generar la retroalimentación automatizada se utilizó la API de un modelo de lenguaje comercial (GPT-4o, OpenAI, 2024), configurado mediante un prompt de sistema diseñado específicamente para esta investigación. El prompt instruía al modelo a comportarse como un tutor de escritura para niños hispanohablantes de entre 9 y 12 años, con las siguientes restricciones: utilizar un registro lingüístico comprensible para la edad, limitar cada respuesta a un máximo de 150 palabras, focalizar los comentarios en no más de dos dimensiones de la rúbrica por entrega, formular al menos una pregunta orientadora al estudiante y evitar corregir el texto directamente (ofrecer pistas en lugar de respuestas).

Antes de la intervención, el equipo investigador realizó un estudio piloto con 40 textos de estudiantes de cuarto grado de otra institución para calibrar el prompt y verificar la adecuación de las respuestas. Se descartaron tres versiones previas del prompt por producir retroalimentación excesivamente técnica o por ofrecer correcciones directas. La versión final fue validada por dos docentes de lengua y literatura que confirmaron la adecuación del nivel lingüístico y la pertinencia didáctica de los comentarios generados.

Procedimiento

La intervención se desarrolló a lo largo de ocho semanas durante el segundo quimestre del año lectivo 2024-2025. En la primera semana se aplicó el pretest: los estudiantes de ambos grupos redactaron un texto narrativo de tema libre con una extensión mínima de 120 palabras, en una sesión de aula de 60 minutos. Los textos fueron digitalizados por el equipo investigador y evaluados con la rúbrica por dos codificadores entrenados.

Durante las semanas 2 a 7, los estudiantes del grupo experimental produjeron un texto semanal (alternando textos narrativos y descriptivos) y recibieron retroalimentación automatizada en un plazo máximo de 24 horas. La retroalimentación se entregó de forma impresa para evitar que el acceso desigual a dispositivos fuera una variable de confusión; cada estudiante recibió una hoja con los comentarios del modelo y un espacio para la reescritura. Los docentes del grupo experimental podían complementar la retroalimentación del modelo con observaciones propias, pero se les pidió que registraran en un diario si lo hacían y en qué aspectos intervenían.

Los estudiantes del grupo de control siguieron el procedimiento habitual de su aula: producción textual semanal con retroalimentación escrita por parte de la docente, sin restricciones sobre su contenido o

CIENCIA INTERDISCIPLINARIA INTERNACIONAL
ISSN: 3121-2948
Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

extensión. En la semana 8 se aplicó el postest con el mismo procedimiento que el pretest, y se realizaron las entrevistas docentes.

Análisis de datos

Los datos cuantitativos se analizaron con el software SPSS v.28. Tras verificar el incumplimiento del supuesto de normalidad en tres de las cinco dimensiones de la rúbrica (prueba de Shapiro-Wilk, $p < .05$), se optó por un enfoque mixto: prueba U de Mann-Whitney para comparaciones entre grupos y prueba de rangos con signo de Wilcoxon para comparaciones intragrupo (pretest-postest). Para las dimensiones que sí cumplieron el supuesto de normalidad, se calcularon pruebas t para muestras independientes como análisis complementario. Los tamaños de efecto se calcularon mediante la d de Cohen para las pruebas paramétricas y la r de Rosenthal ($r = Z / \sqrt{N}$) para las no paramétricas. Se adoptó un nivel de significación de .05 para todas las pruebas.

Los datos cualitativos procedentes de las entrevistas fueron transcritos literalmente y analizados mediante análisis temático reflexivo siguiendo las fases propuestas por Braun y Clarke (2021): familiarización con los datos, generación de códigos iniciales, búsqueda de temas, revisión de temas, definición y denominación, y redacción del informe. La codificación se realizó de forma manual con apoyo del software ATLAS.ti v.23. Para fortalecer la credibilidad del análisis, dos miembros del equipo codificaron de manera independiente cuatro entrevistas, alcanzando un porcentaje de acuerdo del 84 % en los códigos de primer nivel.

Consideraciones éticas

El estudio contó con la aprobación del Comité de Ética de Investigación en Seres Humanos de la Universidad Estatal de Bolívar (Resolución CEISH-UEB-2024-047). Se obtuvo consentimiento informado firmado por los representantes legales de todos los estudiantes participantes y asentimiento verbal de los propios estudiantes. Los textos fueron anonimizados antes de su envío a la API del modelo de lenguaje, eliminando nombres, apellidos y cualquier dato que permitiera identificar al estudiante. Los datos se almacenaron en servidores institucionales con acceso restringido mediante contraseña. La participación fue voluntaria y los representantes fueron informados de su derecho a retirar a sus hijos del estudio en cualquier momento sin consecuencia alguna.

RESULTADOS

Resultados cuantitativos

Antes de la intervención, las puntuaciones del pretest no mostraron diferencias estadísticamente significativas entre el grupo experimental y el grupo de control en ninguna de las cinco dimensiones de la rúbrica (U de Mann-Whitney, $p > .05$ en todos los casos), lo que indica una equivalencia inicial aceptable entre los grupos. La Tabla 1 presenta las puntuaciones descriptivas del pretest y el postest para ambos grupos.

Tabla 1. Puntuaciones descriptivas por dimensión, grupo y momento de medición

Dimensión	GE Pre M(DE)	GE Pos M(DE)	GC Pre M(DE)	GC Pos M(DE)
Coherencia	2.14 (0.71)	3.02 (0.64)	2.08 (0.69)	2.41 (0.72)
Cohesión	1.92 (0.78)	2.81 (0.69)	1.97 (0.74)	2.33 (0.77)
Adec. léxica	2.33 (0.65)	2.72 (0.61)	2.29 (0.68)	2.58 (0.70)
Corr. gramatical	1.88 (0.82)	2.77 (0.70)	1.91 (0.79)	2.22 (0.81)
Estr. textual	2.05 (0.73)	2.61 (0.68)	2.02 (0.71)	2.47 (0.74)

Nota. GE = grupo experimental; GC = grupo de control; Pre = pretest; Pos = postest; M = media; DE = desviación estándar. Escala de 1 a 4 puntos por dimensión.

Las comparaciones postest entre grupos (Tabla 2) revelaron diferencias estadísticamente significativas en tres de las cinco dimensiones. En coherencia, el grupo experimental obtuvo puntuaciones superiores al

CIENCIA INTERDISCIPLINARIA INTERNACIONAL

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

grupo de control (U = 1 421.5, p = .003, r = .26). La dimensión de cohesión también favoreció al grupo experimental (U = 1 512.0, p = .011, r = .22), al igual que la corrección gramatical (U = 1 446.0, p = .007, r = .24). Los tamaños de efecto, interpretados según los umbrales de Cohen (1988), se situaron en el rango mediano.

Las dimensiones de adecuación léxica (U = 1 734.5, p = .132) y estructura textual (U = 1 698.0, p = .089) no alcanzaron el umbral de significación estadística. Si bien ambos grupos mejoraron sus puntuaciones del pretest al postest en estas dimensiones, la magnitud del cambio fue similar.

Tabla 2. Comparaciones postest entre grupos mediante prueba U de Mann-Whitney

Dimensión	U	Z	p	r
Coherencia	1421.5	-2.94	.003*	.26
Cohesión	1512.0	-2.53	.011*	.22
Adec. léxica	1734.5	-1.51	.132	.13
Corr. gramatical	1446.0	-2.71	.007*	.24
Estr. textual	1698.0	-1.70	.089	.15

Nota. * p < .05. r = tamaño del efecto de Rosenthal.

Las comparaciones intragrupo (pretest-postest) mediante la prueba de Wilcoxon confirmaron que el grupo experimental mejoró de manera significativa en las cinco dimensiones (p < .01 en todos los casos), mientras que el grupo de control mostró mejoras significativas solamente en coherencia (p = .028) y adecuación léxica (p = .034). La ganancia media del grupo experimental en la puntuación total de la rúbrica fue de 3.84 puntos (sobre 20), frente a 2.01 puntos en el grupo de control.

Resultados cualitativos

El análisis temático de las entrevistas docentes produjo cuatro temas principales: inmediatez y volumen de la retroalimentación, pertinencia del lenguaje, tensiones con la autonomía pedagógica y condiciones de viabilidad.

Inmediatez y volumen. Diez de los 12 docentes entrevistados destacaron la rapidez con la que el modelo generaba comentarios como la ventaja más perceptible. Una docente de sexto grado señaló: "En mi aula tengo 33 niños. Revisar un texto de cada uno me toma una tarde completa. La herramienta lo hacía en minutos y les daba comentarios a cada uno, no solo una marca roja" (Docente 4, entrevista, semana 8). Varios docentes mencionaron que la inmediatez permitía que los estudiantes revisaran su texto al día siguiente, "cuando todavía se acordaban de lo que querían decir" (Docente 7).

Pertinencia del lenguaje. Las opiniones fueron menos homogéneas en este punto. Siete docentes consideraron que la mayor parte de los comentarios resultaban comprensibles para los estudiantes, pero cinco señalaron que en ocasiones el modelo empleaba términos que los niños no manejaban, como "concordancia temporal" o "conectores causales". Una docente observó que "a veces les decía cosas correctas, pero con palabras que yo no usaría con un niño de quinto" (Docente 2). Dos docentes reportaron haber reescrito manualmente algunos comentarios del modelo antes de entregarlos.

Tensiones con la autonomía pedagógica. Cuatro docentes expresaron incomodidad ante la sensación de que una herramienta externa evaluaba el trabajo de sus estudiantes sin conocer el proceso de aula. "Yo sé que Martín tiene problemas en casa y que el hecho de que haya escrito 10 líneas ya es un logro. La máquina no sabe eso" (Docente 9). No obstante, otros docentes interpretaron la herramienta como un "segundo par de ojos" que les permitía focalizar su propia retroalimentación en aspectos que el modelo no cubría, como la motivación y la creatividad.

CIENCIA INTERDISCIPLINARIA INTERNACIONAL
ISSN: 3121-2948
Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

Condiciones de viabilidad. Todos los docentes coincidieron en que la continuidad del uso dependería de tres factores: estabilidad de la conexión a internet (que en Guaranda presenta interrupciones frecuentes), gratuidad o bajo costo de la herramienta y disponibilidad de formación para su manejo. Tres docentes indicaron que no se sentirían capaces de configurar la herramienta por su cuenta sin asistencia técnica.

DISCUSIÓN

Los resultados de este estudio indican que la retroalimentación automatizada generada por un modelo de lenguaje produjo mejoras significativas en tres dimensiones de la producción escrita (coherencia, cohesión y corrección gramatical) de estudiantes de quinto y sexto grado en escuelas públicas de Guaranda. Estas mejoras, con tamaños de efecto en el rango mediano, coinciden parcialmente con los hallazgos de Dai et al. (2023) en educación superior y con los de Escalante et al. (2023) en educación secundaria, aunque las magnitudes reportadas en esos estudios fueron ligeramente mayores, lo cual puede atribuirse a las diferencias de edad y nivel de competencia escritora de las muestras.

La ausencia de diferencias significativas en adecuación léxica y estructura textual merece atención. Una explicación posible es que estas dos dimensiones implican habilidades de planificación textual y selección de vocabulario que dependen en mayor medida de la instrucción explícita y del modelado docente, y en menor medida de la retroalimentación sobre el borrador ya producido. Graham et al. (2022) han argumentado que la enseñanza de estrategias de planificación previa a la escritura incide más en la estructura global del texto que las observaciones posteriores al borrador, lo que podría explicar por qué una intervención centrada en retroalimentación posescritura no alcanzó a modificar esta dimensión.

Un hallazgo que contradice parcialmente las expectativas iniciales de esta investigación es la magnitud del efecto en corrección gramatical ($r = .24$). Los investigadores anticipaban que esta sería la dimensión con mayor ganancia, dado que la detección de errores gramaticales es una de las capacidades mejor documentadas de los LLM (Kasneci et al., 2023). Sin embargo, el efecto fue comparable al obtenido en coherencia, una dimensión de naturaleza más compleja. Una interpretación plausible es que la retroalimentación del modelo sobre coherencia, al formular preguntas como "¿qué relación tiene esta oración con la anterior?" o "¿podrías explicar cómo se conecta esta idea con tu tema principal?", activó procesos metacognitivos en los estudiantes que tuvieron un impacto proporcionalmente mayor al esperado.

Los resultados cualitativos revelan una tensión que la literatura sobre IA en educación ha señalado con insistencia: la distancia entre la eficacia técnica de la herramienta y su integración pedagógica real (Holmes et al., 2022). Los docentes reconocieron beneficios operativos (inmediatez, volumen), pero varios expresaron reservas que trascienden lo funcional y alcanzan la dimensión profesional: la percepción de que un agente externo evalúa sin conocer el contexto del estudiante. Este hallazgo se alinea con lo reportado por Nazaretsky et al. (2022), quienes encontraron que la confianza de los docentes en los sistemas de IA educativa disminuye cuando perciben que la herramienta opera como una "caja negra" sobre la que no tienen control.

Es necesario reconocer que la intervención incluyó una decisión de diseño que pudo haber mediado los resultados: la retroalimentación se entregó en formato impreso, no en pantalla. Esta decisión, tomada para controlar la desigualdad de acceso a dispositivos, implica que los estudiantes no interactuaron directamente con el modelo en tiempo real, lo cual reduce la posibilidad de diálogo iterativo y limita la intervención a un esquema de retroalimentación unidireccional. Estudios futuros que permitan la interacción directa alumno-modelo podrían generar efectos distintos, tanto positivos (mayor personalización) como negativos (dependencia, distracción).

Otra limitación que exige cautela en la interpretación es la ausencia de seguimiento a mediano plazo. Las mejoras documentadas corresponden al periodo inmediato posterior a la intervención y no permiten afirmar que las ganancias se mantengan una vez retirado el andamiaje automatizado. La investigación

ISSN: 3121-2948
Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

sobre transferencia de habilidades de escritura (Graham, 2022) sugiere que los efectos de las intervenciones de corta duración tienden a diluirse si no se sostienen con prácticas complementarias. Un diseño con medición de seguimiento a tres o seis meses aportaría evidencia más robusta.

La composición de la muestra también limita la generalización de los hallazgos. Participaron 127 estudiantes de dos escuelas urbanas de Guaranda; la población rural de la provincia de Bolívar, con menor acceso a conectividad y mayor presencia del kichwa como lengua de uso cotidiano, no estuvo representada. Extender esta indagación a escuelas rurales y bilingües permitiría evaluar si la herramienta opera de manera equitativa o si, por el contrario, amplía brechas preexistentes, una preocupación recurrente en la literatura sobre equidad e IA (Baker y Hawn, 2022).

A pesar de estas limitaciones, los hallazgos ofrecen implicaciones prácticas concretas para la política educativa provincial y para las instituciones de formación docente del Ecuador. En primer lugar, la retroalimentación automatizada no debería plantearse como un sustituto del docente, sino como un recurso complementario que libera tiempo para que el profesorado atienda los aspectos de la escritura que el modelo no alcanza: la intención comunicativa, la creatividad, la dimensión emocional del texto. En segundo lugar, la implementación de estas herramientas requiere formación específica para los docentes, no solo en su operación técnica, sino en la toma de decisiones sobre cuándo y cómo integrar el feedback del modelo en la secuencia didáctica. Las preocupaciones expresadas por los docentes entrevistados sobre la dependencia tecnológica y el lenguaje del modelo son señales que los programas de capacitación deberían abordar de forma explícita.

CONCLUSIONES

La retroalimentación automatizada generada por un modelo de lenguaje produjo mejoras estadísticamente significativas en coherencia, cohesión y corrección gramatical en la producción escrita de estudiantes de quinto y sexto grado de educación general básica en Guaranda, Ecuador, con tamaños de efecto medianos. Las dimensiones de adecuación léxica y estructura textual no mostraron diferencias atribuibles a la intervención.

Los docentes participantes reconocieron la utilidad operativa de la herramienta, aunque expresaron reservas legítimas sobre su integración en un proceso pedagógico que exige sensibilidad al contexto del estudiante. La tensión entre eficacia técnica y pertinencia contextual constituye un eje que investigaciones posteriores deberán seguir explorando, particularmente en entornos educativos con alta heterogeneidad sociocultural como los de la sierra andina ecuatoriana.

Tres líneas de investigación derivadas de este trabajo merecen atención: evaluar el efecto de la interacción directa y dialogada entre el estudiante y el modelo de lenguaje (frente al esquema unidireccional aquí implementado), medir la sostenibilidad de las ganancias escriturales a mediano plazo y comparar el desempeño de modelos de lenguaje abiertos frente a modelos comerciales en la generación de retroalimentación adaptada a edades tempranas.

REFERENCIAS BIBLIOGRÁFICAS

Baker, R. S. y Hawn, A. (2022). Algorithmic bias in education. International Journal of Artificial Intelligence in Education, 32(4), 1052–1092. https://doi.org/10.1007/s40593-021-00285-9

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. En J. M. Spector, M. D. Merrill, J. Elen y M. J. Bishop (Eds.), Handbook of research on educational communications and technology (4.ª ed., pp. 505–518). Springer. https://doi.org/10.1007/978-1-4614-3185-5_39

Braun, V. y Clarke, V. (2021). Thematic analysis: A practical guide. SAGE Publications.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2.ª ed.). Lawrence Erlbaum Associates.

Crompton, H. y Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.-S., Gašević, D. y Chen, G. (2023). Can large language models provide useful feedback on research papers? A

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

large-scale empirical analysis. arXiv. https://doi.org/10.48550/arXiv.2310.01783

Escalante, J., Hernández, R. y Valdivia, M. (2023). Retroalimentación automatizada mediante ChatGPT en textos argumentativos de educación secundaria. Revista Iberoamericana de Educación, 92(1), 45–63. https://doi.org/10.35362/rie9215741

Flotts, M. P., Manzi, J., Polloni, M. P., Carrasco, M., Zambra, C. y Abarzúa, A. (2021). Aportes para la enseñanza de la escritura a partir de resultados de evaluaciones estandarizadas en América Latina. Revista Latinoamericana de Estudios Educativos, 51(2), 169–198.

Graham, S. (2022). A revised writer(s)-within-community model of writing. Educational Psychologist, 57(2), 85–110. https://doi.org/10.1080/00461520.2021.2001387

Graham, S., Hebert, M. y Harris, K. R. (2022). Formative assessment and writing: A meta-analysis. The Elementary School Journal, 122(4), 525–550. https://doi.org/10.1086/719856

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

Holmes, W., Bialik, M. y Fadel, C. (2022). Artificial intelligence in education: Promises and implications for teaching and learning (2.ª ed.). Center for Curriculum Redesign.

Instituto Nacional de Estadística y Censos [INEC]. (2022). Proyecciones poblacionales a nivel cantonal 2020-2025. INEC.

Instituto Nacional de Evaluación Educativa [INEVAL]. (2023). Resultados nacionales Ser Estudiante 2023: Educación General Básica. INEVAL.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Kruber, 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

Ministerio de Educación del Ecuador. (2021). Currículo priorizado con énfasis en competencias comunicacionales, matemáticas, digitales y socioemocionales: Subnivel Medio de Educación General Básica. Ministerio de Educación.

Nazaretsky, T., Ariely, M., Cukurova, M. y Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. British Journal of Educational Technology, 53(4), 914–931. https://doi.org/10.1111/bjet.13232

OpenAI. (2023). GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2024). GPT-4o system card. OpenAI. https://openai.com/index/gpt-4o-system-card/

Organisation for Economic Co-operation and Development [OECD]. (2021). OECD digital education outlook 2021: Pushing the frontiers with artificial intelligence, blockchain and robots. OECD Publishing. https://doi.org/10.1787/589b283f-en

Peterson, S. S. y McClay, J. K. (2021). Teachers' written comments and student writing in elementary classrooms. Written Communication, 38(3), 391–428. https://doi.org/10.1177/07410883211006360

Sánchez-Prieto, J. C., Trujillo-Torres, J. M., Gómez-García, M. y Gómez-García, G. (2022). Mapping research on artificial intelligence in education in Latin America: A bibliometric analysis. Education Sciences, 12(12), 918. https://doi.org/10.3390/educsci12120918

Shadish, W. R., Cook, T. D. y Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Sotomayor, C., Gómez, G., Jéldrez, E. y Bedwell, P. (2021). Análisis de la retroalimentación escrita de profesores a textos narrativos de estudiantes de educación primaria. Estudios Pedagógicos, 47(1), 281–299. https://doi.org/10.4067/S0718-07052021000100281

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W. y Warschauer, M. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. Learning and Instruction, 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

UNESCO. (2021). AI and education: Guidance for policy-makers. UNESCO Publishing. https://doi.org/10.54675/PCSP7350

UNESCO. (2023). Guidance for generative AI in education and research. UNESCO Publishing. https://doi.org/10.54675/EWZQ9977

Valverde-Berrocoso, J., del Carmen Garrido-Arroyo, M., Burgos-Videla, C. G. y Morales-Cevallos, M. B.

ISSN: 3121-2948

Periodicidad: Trimestral, Volumen: 4, Número: 3, Año: 2026 páginas 70-81

(2022). Trends in educational research about e-learning: A systematic literature review (2009–2018). Sustainability, 12(12), 5153. https://doi.org/10.3390/su12125153

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. y Polosukhin, I. (2017). Attention is all you need. En I. Guyon et al. (Eds.), Advances in Neural Information Processing Systems 30 (pp. 5998–6008). Curran Associates.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.

Warschauer, M. y Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies: An International Journal, 3(1), 22–36. https://doi.org/10.1080/15544800701771580

Wisniewski, B., Zierer, K. y 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. 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

Zheng, L., Niu, J., Zhong, L. y Gyasi, J. F. (2023). The effectiveness of artificial intelligence on learning outcomes and learning persistence in primary and secondary education: A meta-analysis. Computers & Education, 203, 104893. https://doi.org/10.1016/j.compedu.2023.104893

Zawacki-Richter, O., Marín, V. I., Bond, M. y Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

Baidoo-Anu, D. y Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52–62. https://doi.org/10.61969/jai.1337500

Mollick, E. R. y Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. The Wharton School Research Paper. https://doi.org/10.2139/ssrn.4391243

Celik, I. (2023). Towards intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138, 107468. https://doi.org/10.1016/j.chb.2022.107468

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J. y Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity, 2021, 8812542. https://doi.org/10.1155/2021/8812542

Su, J., Zhong, Y. y Ng, D. T. K. (2022). A meta-review of literature on educational approaches for teaching AI at the K-12 levels in the Asia-Pacific region. Computers and Education: Artificial Intelligence, 3, 100065. https://doi.org/10.1016/j.caeai.2022.100065

Kim, N. J. y Kim, M. K. (2022). Teacher's perceptions of using an artificial intelligence-based educational tool for scientific writing. Frontiers in Education, 7, 755914. https://doi.org/10.3389/feduc.2022.755914

Luckin, R. y Cukurova, M. (2019). Designing educational technologies in the age of AI: A learning sciences-driven approach. British Journal of Educational Technology, 50(6), 2824–2838. https://doi.org/10.1111/bjet.12861

Mao, J., Chen, B. y Liu, J. C. (2024). Generative AI and its impact on personalized intelligent tutoring systems. Education and Information Technologies, 29(1), 1443–1464. https://doi.org/10.1007/s10639-024-12580-w